Asymptotic Laws for Joint Content Replication and Delivery in Wireless Networks

S. Gitzenis, *Member, IEEE,* G. S. Paschos, and L. Tassiulas, *Member, IEEE*

***Abstract*—We study the scalability of multihop wireless communications, a major concern in networking, for the case that users access content replicated across the nodes. In contrast to the standard paradigm of randomly selected communicating pairs, content replication is efficient for certain regimes of file popularity, cache and network size. Our study begins with the detailed joint content replication and delivery problem, a hard combinatorial optimization, which is reduced to the simpler replication density problem whose performance is of the same order as the original. Assuming a Zipf popularity law, and letting the number of files and network nodes both go to infinity, we identify scaling capacity laws from $O\left(\sqrt{N}\right)$ down to $O(1)$.**

I. Introduction

A novel technology, key for the future Internet, is content-based networking [1]: in this paradigm, data requests are placed on content, as opposed to network address, and routes are formed based on content provision and user interest. Content caching is a well-known technique that exploits the temporal as well as spatial vicinity of user requests to improve the user-perceived Quality of Service (QoS) and network performance metrics. Its merits have been widely demonstrated in various networking paradigms, such as Content Delivery Networks (CDNs), Publish-Subscribe and Peer-to-Peer (P2P).

Another major technology for future networks is wireless networking, as it supports the exemplar of ubiquitous access for mobile users. Unfortunately, long multihop communications are known not to scale [2] (i.e., the maximum common rate for all flows is inversely proportional to the average number of hops). Considering, moreover, the volatility of wireless communications and the associated performance bottlenecks, it becomes quite important to assess the performance benefits of caching and its effect on the system sustainability.

Aiming to address the performance and sustainability of wireless networks with caching, after reviewing past and related work and the asymptotic notation in Sections II and III, this work makes the following contributions:

1) It formulates the optimization of content replication jointly with routing for minimizing the required wireless link capacity (Section IV), a hard combinatorial problem that looks at the detailed delivery routes and the cache content at each node; this is step-by-step reduced to
2) a simple and mathematically tractable problem, whose scope spans just to the macroscopic frequency of content replication (Section V). An efficient solution is designed, and is shown to be order efficient to the optimal solutions of both the simplified and the original hard problem.
3) Using the above solution, in Section VI, we study and analyze the asymptotic laws of the required link capacity as the network (number of nodes) and content (number of files) both scale to infinity. Link capacity is shown to scale from $O\left(\sqrt{N}\right)$, the non-sustainable regime for the wireless network of [2], down to $O(1)$. The latter implies that, provided sufficiently broad wireless links, the system performance is unaffected of its size.
4) The precise conditions relating the content volume vs. the network size, the cache size and the file popularity distribution are presented for caching to make a difference in the system performance and scalability.

The above are summarized in a discussion in Section VII.

II. Problem Background & Related Work

Consider a wireless network of N nodes, where multihop communications are used to exchange data: in such a network, the maximum throughput per node scales as $O(1/\sqrt{N})$ [2]. This celebrated, albeit pessimistic result essentially states that per-user throughput drops to zero as the network expands. The source of this decline lies in the assumed uniform traffic matrix for the generation of throughput demand: the average communicating pair distance rises as $\Theta\left(\sqrt{N}\right)$.

This result stimulated a series of attempts to overrun it, as in [3], where cooperative transmissions are employed to preserve the user throughput, still hardly avoiding the $1/\sqrt{N}$ law. In [4], non-uniform transfer matrices are analyzed, identifying laws for various types of flows (e.g., asymmetric, multicast, etc.). Even in the case that a multihop wireless network is aided by infrastructure, to overcome the $1/\sqrt{N}$ law, quite many base stations $\Omega\left(\sqrt{N}\right)$ are required [5]. Nonetheless, the limitation is shown to be in nature geometrical [6], and, thus, organically linked to Maxwell's electromagnetic theory.

In this work (and [7], an earlier version), we focus on nodes generating requests on particular *content* (i.e., files), as opposed to specific destinations. This is a major shift, as files may be cached in multiple nodes in the network; hence, requests can be potentially served from multiple locations, close to their origin. Moreover, we assume the standard uniform distribution about the request origin, along with a power law regarding the requested content. The goal is to investigate whether the enhancement of caching can turn sustainable a large wireless network, i.e., overcome the $O(1/\sqrt{N})$ law of [2].

Caching is a technique well-known to improve performance in various domains. In wireless networks, performance benefits arise from the hop reduction [8]; this work revisits and enhances the basic model of [8] in the direction of determining the associated scaling laws. In the Publish-Subscribe paradigm, caching is employed to preserve information spatio-temporally and shield against link breakages and mobility

[9]. In wireless meshes, cooperative caching can improve performance by means of implementation [10].

An exploration of the benefit of caching in the scaling of large networks has been conducted in [11], but on a different perspective than ours. First, [11] uses arbitrary traffic matrix. Second, the cache contents are input parameters, whereas, in this work (and in [8]), replication is a key part of the optimization. Last, [11] assumes the paradigm of cooperative transmissions of [3], which leads to a delivery strategy different than shortest path routing. In total, the network can become approximately sustainable with the use of an hierarchical tree structure of transmissions over arbitrarily long links. However, it should be stressed that the results of [3], [11] depend on the particular signal attenuation parameters assumed.

Here, we study the square grid, a well known 2D model [12] of wireless networks' topology and connectivity, and symmetric traffic about its origin. This approach is less general than [11], aiming to identify definite asymptotic laws and shred light on whether caching can make the system scalable.

To this end, we consider the Zipf law for the non-uniform file requests, as in [8]. Literature provides ample evidence [13], [14] that the file popularity in the Internet follows such a power law. The Zipf parameter ranges from 0.5 [15] to 3 [16] depending on the application: low values are representative in routers, intermediate values in proxies and higher values in mobile applications [17], [18]—see also references therein.

III. Asymptotic Notation

Let us define the notation used in the asymptotic laws that follow. Let f and g be real functions. Then, $f \in o(g)$ if

$$\text{for any } k > 0, \text{ there exists } x_\circ \text{ s. t. for } x \geq x_o, \left|\frac{f(x)}{g(x)}\right| \leq k.$$

Although $o(g)$ defines a set of functions, it is customary to write $f = o(g)$ (slightly abusing notation), instead of $f \in o(g)$.

Moreover, $f = O(g)$, if there exists a $k > 0$ such that $f(x)$ is eventually, in absolute value, less or equal to $kg(x)$, that is

$$\text{there exist } k > 0, x_o > 0 \text{ s. t. for } x \geq x_o, \left|\frac{f(x)}{g(x)}\right| \leq k.$$

Using such a k, we can write that $f \overset{\lim}{\leq} kg$ and $f \overset{\lim}{<} k'g$, if f, g are positive functions and $k' > k$.

If the inequalities on the above definitions are reversed, i.e., $|f(x)/g(x)| \geq k$, we get that $f = \omega(g)$, or $f = \Omega(g)$; using such a k, $f \overset{\lim}{\geq} kg$ and $f \overset{\lim}{>} k'g$, for positive f, g and $k' < k$.

In the case that $f \overset{\lim}{\leq} g$ and $f \overset{\lim}{\geq} g$, we write that $f \sim g$.

Last, $f = \Theta(g)$ if $f = \Omega(g)$ and $f = O(g)$.

An important consequence of the above is that $f = O(g)$ does not imply $f \overset{\lim}{<} g$—e.g., consider $f(x) = 2g(x)$; however, the reverse is true. Moreover, if $f \overset{\lim}{<} g$, then $g - f = \Theta(g)$.

IV. Basic Definitions and the General Problem

Assume N peers, with N being a square of an integer, indexed by $n \in \mathcal{N} \triangleq \{1, 2, \ldots, N\}$, arranged on a square grid on the plane of size $\sqrt{N}$ rows times $\sqrt{N}$ columns. Each node is connected to its four neighbors that lie next to it on the same row or column with non-directed links. By keeping the node density fixed and increasing the network size N, we obtain a scaling network similar to [2]. Moreover, to avoid boundary effects, we consider a toroidal structure as in [19].

This grid structure explicitly considers discrete nodes, in contrast to the average cache capacity density of [8]. Unless nodes coordinate transmissions in complex schemes (e.g., [3], [11]), it is a reasonable communications scheme: the communication range is limited due to attenuation and interference; using a frequency reuse factor appropriate to the physical layer (or TDMA, or random access at the MAC layer), the network layer is abstracted to the lattice [12]. Last, this setup essentially matches the random communicating pairs of [2].

Nodes (or users located therein) generate requests to access files/data, indexed by $m \in \mathcal{M} \triangleq \{1, 2, \ldots, M\}$. Each node n is equipped with a cache/buffer, whose contents are denoted by the set $\mathcal{B}_n$, a subset of $\mathcal{M}$. If a request at node n regards a file m that lies in $\mathcal{B}_n$, then it is served locally. Due to the limited buffer capacity, m will often be not available in $\mathcal{B}_n$, thus, node n will have to request m over the network from some other node w that keeps m in its cache. Let, then, $\mathcal{W}_m \subseteq \mathcal{N}$ be the set of nodes that maintain m in their caches.

Let, moreover, K be the storage capacity of nodes' cache measured in the number of files it can store. This means that all M files are of the same (unit) size, placing a constraint on the cardinality of cache contents $|\mathcal{B}_n| \leq K$. The generalization to variable sized files can be still captured in this framework by splitting each large file into multiple unit segments, and then treating its segments as separate, independent files[1].

For the problem of replication not to be trivial, it should be $K < M$, which implies that each node has to select the files to buffer in its cache (later, in the asymptotic laws, we shift our attention to $K \ll M$). Moreover, for the network to have sufficient memory to store each file at least once, it should be

$$KN \geq M. \tag{1}$$

Last, let each node $n \in \mathcal{N}$ generate requests for data at a rate of λ_n. Although the definitions of Section IV-A cover the general case, we stick to the symmetric node traffic, of uniform request rate across nodes, $\lambda_n = \lambda$. Each request regards a particular file $m \in \mathcal{M}$, depending on the file m's *popularity* p_m. In essence, $[p_m]$ is a probability distribution, i.e., sets the probability of a request for a given file. Clearly, each file's replication should be based on its popularity: e.g., to minimize the network traffic, popular files should be stored densely.

A. General Replication-Routing Problem

Network traffic is the criterion to decide about the file replication, and in particular, the *lowest link capacity* that can sustain the request demand. Defining C_ℓ as the rate of traffic carried by link ℓ, the network is stable, not discarding requests, only if the capacity of link ℓ exceeds C_ℓ. In the primary formulation of the problem, we consider the *worst* link case, i.e., the most loaded link, $\max_\ell C_\ell$. Next, we also consider the *average* traffic over the links, $\operatorname{avg}_\ell C_\ell$. In both

[1]The presented framework allows for independent storage of a file's segments, dispensing with the need to keep them all in the same cache.

Symbol Definition	Symbol	Set of allowed values	Set Cardinality
Node	n	$\mathcal{N}$	$N \triangleq 4^{\nu}$
Alternate Node notation	(x,y)	$\{1,\ldots,N^{\frac{1}{2}}-1\}^2$	N
File/Data	m	$\mathcal{M}$	M
Buffer/Cache contents	$\mathcal{B}_n$	See Table II	
Path (with use probability)	$[r; v_1,\ldots,v_k]$		
Set of Routes $\mathcal{R}_{n,m,i}$ from node n to file m	$\left\{\left[r_i^{n,m}; v_{i1}^{n,m},\ldots,v_{ik_{i,n,m}}^{n,m}\right]\right\}$		
Route $\mathcal{R}_{n,m,i}$ probability	$r_i^{n,m}$	$(0,1]$	
Node maintaining file m in its cache	w_m	$\mathcal{W}_m$	W_m
Node that serves client node n's requests on m	$w_{m,n}$		
Nodes served by node w_m on their requests of m	n	$\mathcal{Q}_{w,m}$	$Q_{w,m}$
Hop count from node n to the serving node(s) $\mathcal{W}_m$	$h(n,m)$	$[0,+\infty)$	
Density of data m	d_m	$\left[\frac{1}{N},1\right]$	
Canonical density of data m	d_m°	$\left\{\frac{1}{N},\frac{4}{N},\ldots,1\right\}$	$1+\nu$
Logarithm of can. density	$\nu_m^{\circ} \triangleq -\log_4 d_m^{\circ}$	$\{0,1,\ldots,\nu\}$	$1+\nu$

TABLE I
LIST OF SYMBOLS USED IN SECTION IV.

Problem:	Worst Link Node Capacity (WN)	Average Link Node Capacity (AN)	Average Link Total Capacity (AT)
Objective	$\min C^{\texttt{WN}}$	$\min C^{\texttt{AN}}$	$\min C^{\texttt{AT}}$
Cost Function	$C^{\texttt{WN}} \triangleq \max_{\ell} C_{\ell}$	$C^{\texttt{AN}} = C^{\texttt{AT}} \triangleq \operatorname{avg}_{\ell} C_{\ell}$	
Optimization Variables	$[\mathcal{B}_n], [\mathcal{R}_{n,m}]$ with $\mathcal{R}_{n,m} = \left\{\left[r_i^{n,m}; v_{i1}^{n,m},\ldots,v_{ik_{i,n,m}}^{n,m}\right]\right\}$		
Constraints on Buffer Contents $[\mathcal{B}_n]$:	For all $n \in \mathcal{N}$, $\lvert\mathcal{B}_n\rvert \leq K$ (2)		$\sum_{n\in\mathcal{N}} \lvert\mathcal{B}_n\rvert \leq KN$ (3)
	For all $n \in \mathcal{N}, \mathcal{B}_n \subseteq \mathcal{M}$ (4)		
	$\bigcup_{n\in\mathcal{N}} \mathcal{B}_n = \mathcal{M}$ (5)		
Constraints on Routes $[\mathcal{R}_{n,m}]$:	For all $n,m,i : n, v_{i1}^{n,m}, v_{i2}^{n,m},\ldots,v_{ik_{i,n,m}}^{n,m}$ is a list of adjacent nodes (6)		
	For all $n,m,i : \mathcal{B}^{n,m}_{v_{ik_{i,n,m}}} \ni m$ (7)		
	For all $n,m : \sum_i r_i^{n,m} = 1$ (8)		
Opt. Value, Arguments	$C^{\texttt{WN}\star}$ $\mathcal{B}_n^{\texttt{WN}\star}, \mathcal{R}_{n,m}^{\texttt{WN}\star}$	$C^{\texttt{AN}\star}$ $\mathcal{B}_n^{\texttt{AN}\star}, \mathcal{R}_{n,m}^{\texttt{AN}\star}$	$C^{\texttt{AT}\star}$ $\mathcal{B}_n^{\texttt{AT}\star}, \mathcal{R}_{n,m}^{\texttt{AT}\star}$

TABLE II
LIST OF JOINT REPLICATION/DELIVERY PROBLEMS.

setups, the replication problem regards finding the cache $\mathcal{B}_n$ at all nodes n, that minimize $\max C_\ell$ or $\operatorname{avg}_\ell C_\ell$, respectively.

The individual node capacity constraint of the primary formulation is expressed through $|B_n| \leq K$, for all $n \in \mathcal{N}$; again, we also consider an alternate version of the capacity constraint, pertaining to the total capacity over the network: $\sum_n |B_n| \leq KN$. This corresponds to a relaxed network setting where the total buffer capacity NK is to be arbitrarily distributed to the nodes. As each file should be stored at least once in the network, in all cases, it has to be $\cup_{n\in\mathcal{N}} B_n = \mathcal{M}$.

Allowing for receiver driven anycast strategy, the network should choose a node $w_{m,n}$ to serve the requests of client n on m; $w_{m,n}$ should be selected from the possibly many candidates of set $\mathcal{W}_m$. Therefore, the replication problem is implicitly linked to a joint *delivery problem* of finding appropriate paths of adjacent nodes $n, v_1, v_2, \ldots, v_k$ from each client n to a node $v_k \in \mathcal{W}_m$ for each file m, that minimize $\max C_\ell$ or $\operatorname{avg}_\ell C_\ell$.

Since there exist multiple routes between a node n and the caches containing m, we should allow splitting the traffic among them, to balance the load among links. Thus, the delivery problem involves finding a set of routes

$$\mathcal{R}_{n,m} = \{\mathcal{R}_{n,m,i}\} = \{[r_i^{n,m}; v_{i1}^{n,m}, v_{i2}^{n,m}, \ldots, v_{ik}^{n,m}]\}$$

from each client node v to a server node $v_{ik}^{n,m} \in \mathcal{W}_m$. On each route $\mathcal{R}_{n,m,i}$, $r_i^{n,m}$ denotes the portion of requests of node n for file m that use path $n, v_{i1}^{n,m}, v_{i2}^{n,m}, \ldots, v_{ik}^{n,m}$ is used by node n. Therefore, it has to be $\sum_i r_i^{n,m} = 1$.

Given the cache contents $[\mathcal{B}_n]$ and routes $[\mathcal{R}_{n,m}]$, it is easy to sum up the traffic C_ℓ per link l; we do not provide the explicit formula, but later we carry out associated computations.

Based on the above, we define three variants of the replication-delivery problem (Table II), beginning with the primary (and hardest) one, relaxed next to simpler versions.

PROBLEM 1 [WORST LINK NODE CAPACITY (WN)]:

Minimize $\max_\ell C_\ell([\mathcal{B}_n], [\mathcal{R}_{n,m}])$, *subject to* (2), (4-8).

PROBLEM 2 [AVERAGE LINK NODE CAPACITY (AN)]:

Minimize $\operatorname{avg}_\ell C_\ell([\mathcal{B}_n], [\mathcal{R}_{n,m}])$, *subject to* (2), (4-8).

PROBLEM 3 [AVERAGE LINK TOTAL CACHE (AT)]:

Minimize $\operatorname{avg}_\ell C_\ell([\mathcal{B}_n], [\mathcal{R}_{n,m}])$, *subject to* (3), (4-8).

As shown in Table II, we use the * notation to denote the optimal value of the objective function, and *a* minimizing pair of the cache contents and routes (as there may exist multiple optimal solutions). Note that in the above definitions, we have omitted from the argument list the parameters $[p_m]$, $[\lambda_n]$ and cache capacity K to contain the notation clutter.

It is clear that these problems are of combinatorial complexity, and thus are not amenable to an easy to compute, closed-form solution. However, as we are interested in asymptotic laws, we will use the simpler structure of the latter problems to design straightforward replication strategies, and compute approximations that are within a constant to the optimal.

First, observe that, as the WN and AN problems share the same constraints, $[\mathcal{B}_n^{\texttt{WN}\star}], [\mathcal{R}_{n,m}^{\texttt{WN}\star}]$ is valid for AN. Moreover, for any $[\mathcal{B}_n], [\mathcal{R}_{n,m}]$, it is $\max_\ell C_\ell \geq \operatorname{avg}_\ell C_\ell$, therefore,

$$C^{\texttt{WN}\star} \geq \operatorname{avg}_{\ell} C_\ell([\mathcal{B}_n^{\texttt{WN}\star}], [\mathcal{R}_{n,m}^{\texttt{WN}\star}]) \geq C^{\texttt{AN}\star}.$$

Second, the AT problem has relaxed constraints in comparison to AN, i.e., any $([\mathcal{B}_n, \mathcal{R}_{n,m}])$ of AN satisfies AT, too, and both problems share a common objective function. Thus,

$$C^{\texttt{AN}\star} = C^{\texttt{AT}}([\mathcal{B}_n^{\texttt{AN}\star}], [\mathcal{R}_{n,m}^{\texttt{AN}\star}]) \geq C^{\texttt{AT}\star}.$$

Combining the above, it follows that

LEMMA 1 [WN VS. AN VS. AT]: $C^{\texttt{WN}\star} \geq C^{\texttt{AN}\star} \geq C^{\texttt{AT}\star}$.

Observe that shortest paths are sufficient in the AN and AT problems. Indeed, consider for some n, m a route $[r_i^{n,m}; v_{i1}^{n,m}, v_{i2}^{n,m}, \ldots, v_{ik_{i,n,m}}^{n,m}]$ in set $\mathcal{R}_{n,m}$ that involves more hops than path $[n, v_1, v_2, \ldots, v_k]$ with $m \in \mathcal{B}_{v_k}$. Replacing it by $[r_i^{n,m}; v_1, v_2, \ldots, v_k]\}$ (or, if the shorter path is already in $\mathcal{R}_{n,m}$, adding the routing probabilities) reduces the average link load thanks to the new path's fewer hops:

LEMMA 2 [AN-AT SHORTEST PATH OPTIMALITY]: *The optimal routes $\mathcal{R}_{n,m}^{\mathtt{AN}\star}, \mathcal{R}_{n,m}^{\mathtt{AT}\star}$ consist of shortest paths only.*

As there may exist multiple paths from a source n to file m with the same hop count; in this case, we are free to arbitrarily distribute traffic among them. Let, therefore, $h(n, m)$ denote the hop-count of a shortest path between n and m. It is

$$\sum_{\ell} C_\ell = \sum_{n \in \mathcal{N}} \sum_{m \in \mathcal{M}} h(n, m) p_m. \tag{9}$$

For simplicity, the above assumes a unit request rate per node $\lambda = 1$; it is clear that the link capacities of WN, AN and AT are directly linear to the arrival rate λ. Then, both sides of (9) sum the total load on the network: the LHS expresses the total load as the sum over all links, while the RHS expresses it as the load generated by each file request of each node.

V. DENSITY-BASED FORMULATIONS

The above formulation takes a *microscopic* view on the precise routes and specific cache contents. Now, we switch to a *macroscopic* view, that considers the frequency of occurrence of each file in the caches. This leads to an easy-to-solve problem, which permits finding the asymptotic link rate.

A. Replication Density and Distance Approximation

Focusing on the frequency of occurrence of each file m in the caches $\mathcal{B}_n$, we define the *replication density* d_m as the fraction of nodes that store file m in the network:

$$d_m = \frac{1}{N} \sum_{n \in \mathcal{N}} \mathbb{1}_{\{m \in \mathcal{B}_n\}}. \tag{10}$$

Under shortest path routing, the inverse of replication density d_m may be regarded in a fluid approximation as the number of peers served by the node that maintains m in its cache; hence, it also represents the size of the area served by a specific location as the source of information m.

Note that in the AT and previous problems, d_m takes values in the set of $\{1/N, 2/N, \ldots, 1\}$, and moreover, that

$$\sum_{m \in \mathcal{M}} d_m = \frac{1}{N} \sum_{n \in \mathcal{N}} |\mathcal{B}_n| \leq \frac{1}{N} \sum_{n \in \mathcal{N}} K = K.$$

Consider now densities $[d_m^{\mathtt{AT}\star}]$ defined from the optimal cache $[\mathcal{B}_n^{\mathtt{AT}\star}]$ and (10). The following is a very useful result (see Appendix A for the proof of all results of this Section) that naturally lead us to a new problem based on the file densities:

LEMMA 3 [AT CAPACITY LOWER BOUND]:

$$C^{\mathtt{AT}\star} \geq \frac{\sqrt{2}}{6} \sum_{m \in \mathcal{M}} \left(\sqrt{\frac{1}{d_m^{\mathtt{AT}\star}}} - 1 \right) p_m.$$

PROBLEM 4 [CONTINUOUS DENSITY (CD)]: *Minimize*

$$C^{\mathtt{CD}}([d_m]) \triangleq \frac{\sqrt{2}}{6} \sum_{m \in \mathcal{M}} \left[\frac{1}{\sqrt{d_m}} - 1 \right] p_m,$$

with respect to $[d_m]$, subject to:

1) *For any $m \in \mathcal{M}$, $\frac{1}{N} \leq d_m \leq 1$,*
2) $\sum_{m \in \mathcal{M}} d_m \leq K$.

In the CD Problem, the optimization variables are the densities d_m, which express the fraction of caches containing file m. In the objective, $d_m^{-\frac{1}{2}} - 1$ approximates the average hop count from a random node to a cache containing m. Weighted by the probability p_m of requests on m, the summation expresses the average link load per request.

It is clear that any $[\mathcal{B}_n]$ that satisfies the WN, AN or AT problems' constraints yields a distribution $[d_m]$ that satisfies the constrains of CD. Combining this with Lemma 1,

THEOREM 4 [CD BOUND]: $C^{\mathtt{CD}\star} \leq C^{\mathtt{AT}\star} \leq C^{\mathtt{AN}\star} \leq C^{\mathtt{WN}\star}$.

The $1/N \leq d_m \leq 1$ are an important difference from the model of [8], which, as seen next, affects the solution, and has a major impact in the asymptotics.

B. CD Problem Solution

First, let us establish the uniqueness of the solution:

LEMMA 5 [CD CONVEXITY]: *CD is a strictly convex optimization problem, and has a unique optimal solution.*

The computation of the solution is relatively easy, using the Karush-Kuhn-Tucker (KTT) conditions. Except for the trivial case of $M > KN$, the summation constraint is satisfied as an equality. Regarding the pair of constraints on each d_m, either one of them can be an equality, or none. This partitions $\mathcal{M}$ into three subsets, one containing files of unit replication density (i.e., stored at every node) $\mathcal{M}_{\uparrow} = \{m : d_m = 1\}$, one containing files stored in just one node $\mathcal{M}_{\downarrow} = \{m : d_m = \frac{1}{N}\}$, and the complementary $\mathcal{M}_{\updownarrow} = \mathcal{M} \setminus (\mathcal{M}_{\uparrow} \cup \mathcal{M}_{\downarrow})$.

If the files are ordered, then the three subsets are ordered:

LEMMA 6 [MONOTONICITY OF SETS $\mathcal{M}_{\uparrow}, \mathcal{M}_{\updownarrow}, \mathcal{M}_{\downarrow}$]: *Provided that p_m is non-increasing on m, the optimal solution for the CD problem takes the form of $\mathcal{M}_{\uparrow} = \{1, 2, \ldots, l-1\}$, $\mathcal{M}_{\updownarrow} = \{l, l+1, \ldots, r-1\}$, and $\mathcal{M}_{\downarrow} = \{r, r+2, \ldots, M\}$, where l and r are integers with $1 \leq l \leq r \leq M+1$.*

The uniqueness of the optimal solution $[d_m^{\mathtt{CD}*}]$ implies the uniqueness of the indices and the three partinitions. We, then, denote by $l^\star, r^\star$, and $\mathcal{M}_{\uparrow}^\star$, $\mathcal{M}_{\updownarrow}^\star$, $\mathcal{M}_{\downarrow}^\star$ their optimal values. Given these, the solution $d_m^{\mathtt{CD}\star}$ takes the form of

$$d_m^{\mathtt{CD}\star} = \begin{cases} 1, & m \in \mathcal{M}_{\uparrow}^\star, \quad \text{(11a)} \\ \dfrac{K - l^\star + 1 - \frac{M - r^\star + 1}{N}}{\sum_{j=l^\star}^{r^\star - 1} p_j^{\frac{2}{3}}} \, p_m^{\frac{2}{3}}, & m \in \mathcal{M}_{\updownarrow}^\star, \quad \text{(11b)} \\ \dfrac{1}{N}, & m \in \mathcal{M}_{\downarrow}^\star. \quad \text{(11c)} \end{cases}$$

Fig. 1 illustrates such an example solution, depicting the density $d_m^{\mathtt{CD}\star}$, indices $l^\star$ and $r^\star$, as well as sets $\mathcal{M}_{\uparrow}^\star$, $\mathcal{M}_{\updownarrow}^\star$, $\mathcal{M}_{\downarrow}^\star$ when file popularities follow the Zipf law (see Section VI-A).

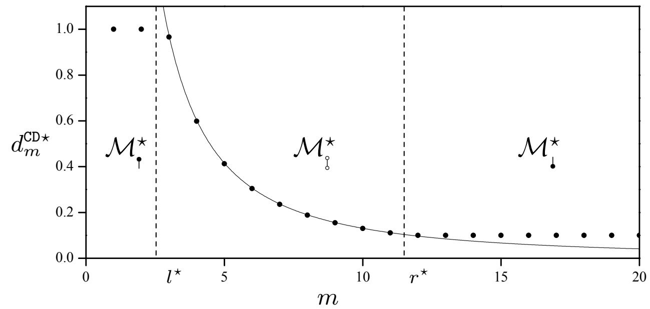

Fig. 1. An example case of density d_m and the $\mathcal{M}^\star_{\uparrow}, \mathcal{M}^\star_{\updownarrow}$ and $\mathcal{M}^\star_{\downarrow}$ partitions. Solid line plots the $\sim m^{-\frac{2\tau}{3}}$ law of $m \in \mathcal{M}^\star_{\updownarrow}$, when p_m follows Zipf's law.

C. Discrete Density Formulation

Now, we restrict our attention to networks with a number of nodes equal to a power of 4, $N = 4^\nu$, and consider a constrained version of the CD Problem of discrete densities:

PROBLEM 5 [DISCRETE DENSITY (DD)]: *Minimize* $C^{\mathtt{DD}}([d_m]) \triangleq C^{\mathtt{CD}}([d_m])$ *with respect to* $[d_m]$*, subject to:*

1) *For any* $m \in \mathcal{M}$*,* $d_m = 4^{-\nu_m}$*, with* $\nu_m \in \{0, 1, \ldots, \nu\}$*.*
2) $\sum_{m \in \mathcal{M}} d_m \leq K$

It is clear that the CD Problem is a relaxed version of the DD Problem, i.e., any $[d_m]$ satisfying DD constraints, satisfies CD's constraints as well, hence $C^{\mathtt{CD}\star} \leq C^{\mathtt{DD}\star}$.

DD, defined on a discrete set, is not easy to solve. Still, it is straightforward to construct an efficient solution $d_m^\circ \triangleq 4^{-\nu_m^\circ}$ setting it to the largest power less or equal to $[d_m^{\mathtt{CD}\star}]$

$$d_m^\circ \triangleq \max\left\{4^{-i} \colon 4^{-i} \leq d_m^{\mathtt{CD}\star},\ i \in \{0, 1, \ldots, \nu\}\right\}. \tag{12}$$

Let, then, $C^{\mathtt{DD}\circ} = C^{\mathtt{CD}\circ}$ be the required link rate of the DD and CD problems for the $[d_m^\circ]$ replication density. Although not optimal, $[d_m^\circ]$ is quite efficient a solution:

THEOREM 7 [CANONICAL PLACEMENT EFFICIENCY]:

$$C^{\mathtt{CD}\star} \leq C^{\mathtt{CD}\circ} < 2C^{\mathtt{CD}\star} + \frac{\sqrt{2}}{6}.$$

D. Replication Policy Design

We present a replication algorithm that allocates the files $\mathcal{M}$ in the caches $[\mathcal{B}_n]$ given the replication densities d_m°. In the algorithm, we represent each node n by its coordinates (x, y) with x and y taking values in $\{0, 1, \ldots, \sqrt{N} - 1\}$. First, we partition the data into the sets $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_\nu$, such that $\mathcal{M}_i$ contains data m of $d_m^\circ = 4^{-i}$. Then, each file m gets assigned to a unique node in the $2^{\nu_m^\circ} \times 2^{\nu_m^\circ}$ submatrix.

Indeed, we start (step 1) by assigning to $\mathcal{B}_{(0,0)}$, the first 1×1 submatrix the elements of $\mathcal{M}_0$, i.e., the files to be replicated at all nodes of the network. Then, in the loop of step 2, we scan all sets $\mathcal{M}_k$ and allocate its elements to the buffers.

First, we repeat the $2^{k-1} \times 2^{k-1}$ first submatrix created so far to the $2^k \times 2^k$ submatrix (loop of step 4), and then, we iteratively take out of set $\mathcal{M}_k$ files at decreasing popularity, filling one node in the $2^k \times 2^k$ submatrix (step 8). The node where each data is placed is selected as a node with the least number of files so far and in case of tie, by scanning the diagonals (step 12). In this way, we spread files to balance the load over the links, a key property for the WN problem.

Algorithm 1 [Cache Data Filling]:

1: $\mathcal{B}_{(0,0)} \leftarrow \mathcal{M}_0$
2: **for** $k \in \{1, 2, \ldots, \nu\}$ **do**
3:   **for** $x = 0, 1, \ldots, 2^k - 1$ **do**
4:     **for** $y = 0, 1, \ldots, 2^k - 1$ **do**
5:       $\mathcal{B}_{(x, y+2^{k-1})} \leftarrow \mathcal{B}_{(x,y)}$ {Replicate cache contents}
6:       $\mathcal{B}_{(x+2^{k-1}, y)} \leftarrow \mathcal{B}_{(x,y)}$
7:       $\mathcal{B}_{(x+2^{k-1}, y+2^{k-1})} \leftarrow \mathcal{B}_{(x,y)}$
8:   **while** $\mathcal{M}_k$ is not empty **do**
9:     $m \leftarrow \arg\min_{m \in \mathcal{M}_k} p_m$. {Find the file m of highest p_m}
10:     $\mathcal{M}_k \leftarrow \mathcal{M}_k \setminus \{m\}$. {Remove it from set $\mathcal{M}_k$}
11:     $\mathcal{S} \leftarrow \arg\min_{(x,y) \in \{0,1,\ldots,2^k-1\}^2} |\mathcal{B}_{(x,y)}|$. {Find the **set** of nodes (x, y) with the minimum number of elements}
12:     Select node (x, y) from $\mathcal{S}$ as the first node by scanning from top left to bottom right the main diagonal, then next diagonal, etc with wrap around in the $2^k \times 2^k$ submatrix. {See Fig. 2}

Denote by $\mathcal{B}_n^\circ = \mathcal{B}_{(x,y)}^\circ$ the cache contents of node (x, y) at the end of the Algorithm 1. Clearly, in $\mathcal{B}_n^\circ$ each data m is replicated $4^{\nu_m^\circ} = N/d_m^\circ$ times in the network, with replicas located every $2^{\nu_m^\circ}$ nodes along the two axis (as in Fig. 3). Next, we study the validity and optimality of Algorithm 1.

E. Validity and Optimality for AN and WN Problem

LEMMA 8 [ALGORITHM 1 CACHE SIZE]: *Algorithm 1 results in caches* $\mathcal{B}_n$ *filled with at most* K *files.*

The above result shows that Algorithm 1 constructs a solution compatible to the WN's and AN's constraints. Let us specify a set of shortest paths $[\mathcal{R}_{n,m}^\circ]$ to use for delivery in conjunction to the caches $[\mathcal{B}_n^\circ]$ (see Fig. 6 in Appendix A):

- In the case that there are multiple w_m nodes at the same hop-count from node n, requests of n about m will be directed to the w_m node at the north and west of node n.
- If n and w_m are on the same row or column of the network, we use the single I-shared shortest path.
- If n and w_m are on different rows and columns of the network, we use equally the two L-shaped shortest paths.

The above routes are optimal for the AN problem, but not necessarily for WN. Define $C^{\mathtt{AN}\circ} \triangleq C^{\mathtt{AN}}([\mathcal{B}_n^\circ], [\mathcal{R}_{n,m}^\circ])$, $C^{\mathtt{WN}\circ} \triangleq C^{\mathtt{WN}}([\mathcal{B}_n^\circ], [\mathcal{R}_{n,m}^\circ])$ and

$$A_{i,j} \triangleq \begin{cases} \frac{1}{K - i - \frac{j}{N}} \sum_{k=i+1}^{M-j} p_k^{2/3}, & \text{if } K - i - \frac{j}{N} > 0, \\ 1, & \text{if } K - i - \frac{j}{N} = 0. \end{cases} \tag{13}$$

Next results establish the order optimality of the constructed solution for the AN and WN problems:

THEOREM 9 [ALGORITHM 1 OPTIMALITY ON AN]:

$$C^{\mathtt{AN}\star} \leq C^{\mathtt{AN}\circ} \leq \frac{1}{2} + \frac{3}{2}\sqrt{2}C^{\mathtt{AN}\star}.$$

$$\begin{bmatrix}
1 & (2^k{-}1)2^k{+}1 & (2^k{-}2)2^k{+}2 & \cdots & 2{\cdot}2^k{-}1 & 2{\cdot}2^k \\
2^k{+}1 & 2 & (2^k{-}2)2^k{+}2 & \cdots & 3{\cdot}2^k{-}1 & 3{\cdot}2^k \\
2{\cdot}2^k{+}1 & 2^k{+}2 & 3 & \cdots & 4{\cdot}2^k{-}1 & 4{\cdot}2^k \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
(2^k{-}2)2^k{+}1 & (2^k{-}3)2^k{+}2 & (2^k{-}4)2^k{+}3 & \cdots & 2^k{-}1 & 4^k \\
(2^k{-}1)2^k{+}1 & (2^k{-}2)2^k{+}2 & (2^k{-}3)2^k{+}3 & \cdots & 2{\cdot}2^k{-}1 & 2^k
\end{bmatrix}$$

Fig. 2. The order of precedence in the $2^k \times 2^k$ matrix (Algorithm 1, step 12).

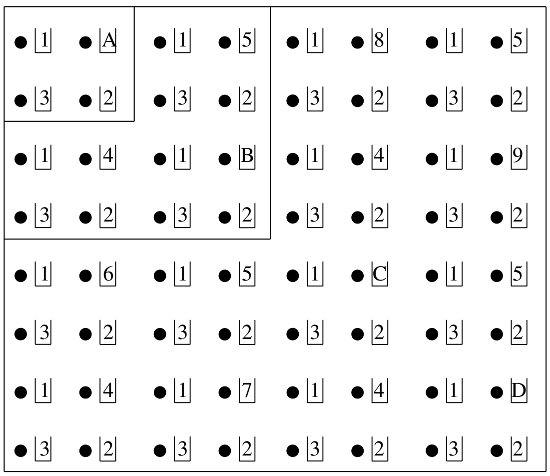

Fig. 3. An example run of Algorithm 1 for $\mathcal{M}_1 = \{1,2,3\}, \mathcal{M}_2 = \{4,5\}, \mathcal{M}_3 = \{6,7,8,9,\mathrm{A},\mathrm{B},\mathrm{C},\mathrm{D},\dots\}$: Files $\{1,\dots,\mathrm{D}\}$ have been placed in the buffers; in subsequent steps, buffers will start receiving their second file. The boxes mark the $2^k \times 2^k$ submatrices.

THEOREM 10 [ALGORITHM 1 OPTIMALITY ON WN]: *The maximum link load $C^{\mathtt{WN}\circ}$ is within a multiplicative constant to the optimal $C^{\mathtt{WN}\star}$ plus an additive term $2 + A_{0,0}/4$, that depends on the distribution $[p_m]$, and cache capacity K.*

VI. ASYMPTOTIC LAWS FOR ZIPF POPULARITY

To study the scaling of link rate, we switch from the arbitrary popularity to the Zipf law, a distribution known to model well various types of network traffic.

A. Zipf Law and Approximations

The Zipf distribution is defined as follows:

$$p_m = \frac{1}{H_\tau(M)} m^{-\tau}, \tag{14}$$

where τ is the power law parameter, indicating the rate of popularity decline as m increases, and $H_\tau(n) \triangleq \sum_{j=1}^{n} j^{-\tau}$ is the truncated (at n) zeta function evaluated at τ (also called the n^{th} *τ-order generalized harmonic number*). The limit $H_\tau \triangleq \lim_{n\to\infty} H_\tau(n)$ is the Riemann zeta function, which converges when $\tau > 1$. We derive an approximation for $H_\tau(n)$ by bounding the sum: for $n \geq m \geq 0$,

$$\int_m^n (x+1)^{-\tau} dx \leq H_\tau(n) - H_\tau(m) \leq 1 + \int_{m+1}^n x^{-\tau} dx, \Rightarrow$$

$$\begin{cases} \frac{(n+1)^{1-\tau}-(m+1)^{1-\tau}}{1-\tau} \leq H_\tau(n) - H_\tau(m) \leq \frac{n^{1-\tau}-(m+1)^{1-\tau}}{1-\tau} + 1, & \text{if } \tau \neq 1, \\ \ln \frac{n+1}{m+1} \leq H_\tau(n) - H_\tau(m) \leq \ln \frac{n+1}{m+2}, & \text{if } \tau = 1. \end{cases} \tag{15}$$

As we are interested in the asymptotic scaling of link rates, for notational simplicity, we remove the multiplicative factor from the objective function of the CD Problem, and refer to the resulting quantity as C. Substituting the solution (11) and plugging in the Zipf distribution into (16), it follows that

$$C \triangleq \sum_{m\in\mathcal{M}} \left(d_m^{-\frac{1}{2}} - 1\right) p_m = C_{\updownarrow} + C_{\downarrow} - \sum_{j=l}^{M} p_m, \tag{16}$$

where $\sum_{j=l}^{M} p_m = O(1)$ (as it lies always in $[0,1]$), and

$$C_{\updownarrow} \triangleq \sum_{m\in\mathcal{M}_{\updownarrow}} \frac{p_m}{\sqrt{d_m}} \overset{(14)}{=} \frac{\left[H_{\frac{2\tau}{3}}(r-1) - H_{\frac{2\tau}{3}}(l-1)\right]^{\frac{3}{2}}}{\sqrt{K_{\updownarrow}}\; H_\tau(M)}, \tag{17}$$

$$C_{\downarrow} \triangleq \sum_{m\in\mathcal{M}_{\downarrow}} \frac{p_m}{\sqrt{d_m}} \overset{(14)}{=} \sqrt{N}\; \frac{H_\tau(M) - H_\tau(r-1)}{H_\tau(M)}, \tag{18}$$

$$K_{\updownarrow} \triangleq \frac{(K-l+1)N - (M-r+1)}{N}. \tag{19}$$

Last, for simplicity, we drop the star notation and use d_m, l and r to refer to the values of the optimal solution.

B. Estimation of l and r

As indices l, r are not provided in a closed form, we derive approximations in order to find the scaling of C.

1) Estimation of l: first, $l \leq K+1$ (hence $l = \Theta(1)$), as $l-1$ represents the number of files cached in all nodes. If $\mathcal{M}_{\updownarrow}$ and $\mathcal{M}_{\downarrow}$ are not empty, using (11b), $d_l < 1$ is equivalent to

$$K - l + 1 - \frac{M-r+1}{N} < l^{\frac{2\tau}{3}} \left[H_{\frac{2\tau}{3}}(r-1) - H_{\frac{2\tau}{3}}(l-1)\right]. \tag{20}$$

If, moreover, the first set $\mathcal{M}_{\downarrow}$ is not empty, i.e., $l > 1$, then $d_{l-1} = 1$. This means that if we attempted to decrease index l by 1, this would violate the density constraints, and result in (11b) a number greater than 1 for d_{l-1}:

$$K - l - \frac{M-r+1}{N} \geq (l-1)^{\frac{2\tau}{3}} \left[H_{\frac{2\tau}{3}}(r-1) - H_{\frac{2\tau}{3}}(l-2)\right]. \tag{21}$$

Thus, provided $l > 1$, it can be uniquely determined as the lowest integer that satififies (20)-(21), which is unique (Theorem 5). An approximation for l can be computed treating (20) as an approximate equality when $\mathcal{M}_{\updownarrow} \neq \emptyset$, or equivalently when $l < r$ (as $d_{l-1} = 1$ and $d_l < 1$):

$$K - l + 1 - \frac{M-r+1}{N} \cong l^{\frac{2\tau}{3}} \left[H_{\frac{2\tau}{3}}(r-1) - H_{\frac{2\tau}{3}}(l-1)\right]. \tag{22}$$

2) Estimation of r: If $\mathcal{M}_{\updownarrow} \cup \mathcal{M}_{\uparrow}$ is not empty, $d_{r-1} > \frac{1}{N} \Leftrightarrow$

$$(K-l+1)N - M + r - 1 > (r-1)^{\frac{2\tau}{3}} \left[H_{\frac{2\tau}{3}}(r-1) - H_{\frac{2\tau}{3}}(l-1)\right]. \tag{23}$$

Again, if set $\mathcal{M}_{\downarrow}$ is not empty, i.e., $r \leq M$, then $d_r = N^{-1}$. Thus, if we attempted increasing index r by one, (11b) would violate the constraint resulting in a density less than N^{-1}:

$$(K-l+1)N - M + r \leq r^{\frac{2\tau}{3}} \left[H_{\frac{2\tau}{3}}(r) - H_{\frac{2\tau}{3}}(l-1)\right]. \tag{24}$$

As before, (23) is an approximate equality if $l < r$, i.e.,

$$(K-l+1)N-M+r-1 \cong (r-1)^{\frac{2\tau}{3}}\left[H_{\frac{2\tau}{3}}(r-1)-H_{\frac{2\tau}{3}}(l-1)\right]. \tag{25}$$

3) Estimation of $^{l}/_{r}$: For all l, r, it is $N > \frac{d_l}{d_{r-1}} = \left(\frac{r-1}{l}\right)^{\frac{2\tau}{3}}$. As before, whenever l and r are not equal to the extremes, i.e., $1 < l < r < M+1$, it holds that $d_{l-1}/d_r = N$. Thus,

$$l \cong rN^{-\frac{3}{2\tau}}. \tag{26}$$

Next, we show that it is impossible to have files cached in *all* nodes unless the popularity parameter τ is greater than $\frac{3}{2}$ (the proofs of the results that follow can be found in Appendix B):

Lemma 11: *If $\tau \leq \frac{3}{2}$, then $l \to 1$.*

We proceed to the asymptotic behavior of the system on the rate C, and indices l and r, which govern partiting to $\mathcal{M}_{\uparrow}$, $\mathcal{M}_{\updownarrow}$ and $\mathcal{M}_{\downarrow}$. The scaling behavior regards the case of the number of nodes N and the number of files M increasing to infinity. We use $\hat{l}$ and $\hat{r}$ to refer to the limits of l and r.

First, a set of basic results establish the upper bound of $C = O\left(\sqrt{N}\right)$, the Gupta-Kumar rate [2]. This is intuitive: if replication is ineffective (e.g., due to large number of files), then the system and its performance essentially reduce to [2].

Lemma 12 [Bound on $C_{\updownarrow}$]: $C_{\updownarrow} = O\left(\sqrt{N}\right)$.

Lemma 13 [Bounds on $C_{\downarrow}$]: $C_{\downarrow} = O\left(\sqrt{N}\right)$. *Furthermore,*

1) *for $\tau < 1$, and $r \overset{\lim}{<} M$, it is $C_{\downarrow} = \Theta\left(\sqrt{N}\right)$,*
2) *for $\tau > 1$, it is $C_{\downarrow} = \Theta\left(\sqrt{N}\left(r^{1-\tau} - M^{1-\tau}\right)\right)$.*
   If, moreover, $r \overset{\lim}{<} M$, then $C_{\downarrow} = \Theta\left(\frac{\sqrt{N}}{r^{\tau-1}}\right)$.

Corollary 14 [Bound on C]: $C = O\left(\sqrt{N}\right)$.

Next, we start the asymptotic analysis, partitioning the space of M, N parameters according to whether they produce singly replicated files or not, i.e., non-empty or empty $\mathcal{M}_{\downarrow}$.

C. Almost Empty $\mathcal{M}_{\downarrow}$

The first case to consider is when the number of nodes N and files M increase towards infinity, and at the same time $\mathcal{M}_{\downarrow}$ remains an almost empty set. We define formally $\mathcal{M}_{\downarrow} \approx \emptyset$ iff $|\mathcal{M}_{\downarrow}| = o(M)$, i.e., the number of its elements is lower order than the total files. For this to happen, M should increase at a slow pace with N, so that the constraint $d_m \geq N^{-1}$ is satisfied for almost all (i.e., up to $o(M)$) files. The extreme case of this is to first let $N \to \infty$, and then have $M \to \infty$, i.e., split the limit of N, M jointly going to infinity to a *double* limit.

To study the asymptotics of C, we first estimate l and r. The almost empty $\mathcal{M}_{\downarrow}$ implies that $M - r = o(M)$, thus $r \sim M$.

Theorem 15 [$\hat{l}$ for Almost Empty $\mathcal{M}_{\downarrow}$]:

1) *For $\tau \leq {}^{3}/_{2}$, $l \to 1$.*
2) *For $\tau > {}^{3}/_{2}$, l converges to the integer solution of*

$$\begin{cases} (K-l+1)l^{-\frac{2\tau}{3}} & < \quad H_{\frac{2\tau}{3}} - H_{\frac{2\tau}{3}}(l-1), \\ (K-l)(l-1)^{-\frac{2\tau}{3}} & \geq \quad H_{\frac{2\tau}{3}} - H_{\frac{2\tau}{3}}(l-2), \end{cases} \tag{27}$$

if such exists and is greater than 1, or 1 otherwise.

Note that an approximate solution of (27) is obtained from

$$K-(l-1) \cong (l-1)\left[H_{\frac{2\tau}{3}} - H_{\frac{2\tau}{3}}(l-1)\right] \overset{(15)}{\cong} 3\frac{l-1}{2\tau-3} \Leftrightarrow$$
$$l \cong 1 + \frac{2\tau-3}{2\tau}K. \tag{28}$$

Next, we study the conditions to have $\mathcal{M}_{\downarrow}$ almost empty:

Theorem 16 [$\mathcal{M}_{\downarrow}$ Almost Empty]: $M - r = o(M)$ *iff*

- *for $\tau < {}^{3}/_{2}$,* $\quad M \overset{\lim}{\leq} \left(1 - \frac{2\tau}{3}\right)KN$,
- *for $\tau = {}^{3}/_{2}$,* $M \ln M \overset{\lim}{\leq} KN$,
- *for $\tau > {}^{3}/_{2}$,* $\quad M \overset{\lim}{\leq} \left[\frac{(K-\hat{l}+1)\left(\frac{2\tau}{3}-1\right)}{\hat{l}^{1-\frac{2\tau}{3}}}\right]^{\frac{3}{2\tau}} N^{\frac{3}{2\tau}}$.

where $\hat{l} = \hat{l}_{\{\tau > \frac{3}{2}, \mathcal{M}_{\downarrow} = \emptyset\}}$ from Theorem 15. If the above inequalities are strict, then $r = M+1$ (and thus $\mathcal{M}_{\downarrow} = \emptyset$).

Theorem 17 [Capacity for Almost Empty $\mathcal{M}_{\downarrow}$]:

- *If $\tau < 1$,* $\quad C = \Theta\left(\sqrt{M}\right)$.
- *If $\tau = 1$,* $\quad C = \Theta\left(\frac{\sqrt{M}}{\log M}\right)$.
- *If $1 < \tau < {}^{3}/_{2}$,* $C = \Theta\left(M^{3/2-\tau}\right)$.
- *If $\tau = {}^{3}/_{2}$,* $\quad C = \Theta\left(\log^{\frac{3}{2}} M\right)$.
- *If $\tau > {}^{3}/_{2}$,* $\quad C = \Theta(1)$.

D. Non-empty $\mathcal{M}_{\downarrow}$

When $M_{\downarrow}$ is non-empty, it is $C_{\downarrow} > 0$. As Corollary 14 shows, C is $O\left(\sqrt{N}\right)$, i.e., the Gupta-Kumar rate [2]. Thus, we turn our attention to identifying the cases that $C = o\left(\sqrt{N}\right)$.

Theorem 18 [$\hat{l}$ and $\hat{r}$ for Non-Empty $\mathcal{M}_{\downarrow}$]: *If M exceeds the condition of Theorem 15,*

- *if $KN - M = \omega(1)$, then we discern the following cases:*

$$\tau < {}^{3}/_{2}: \quad l \to 1, \quad r \sim \frac{3-2\tau}{2\tau}(KN-M), \tag{29}$$
$$\tau = {}^{3}/_{2}: \quad l \to 1, \quad r \ln r \sim KN - M, \tag{30}$$

$\tau > {}^{3}/_{2}$ *and* $M \overset{\lim}{\leq} (K-\beta)N$:

$$l \to \hat{l} \cong \alpha\left[K + 1 - \lim \frac{M}{N}\right], \tag{31}$$
$$r \sim \alpha\left[KN^{\frac{3}{2\tau}} - \frac{M}{N^{1-\frac{3}{2\tau}}}\right]. \tag{32}$$

$\tau > {}^{3}/_{2}$ *and* $M \overset{\lim}{>} (K-\beta)N$:

$$l \to 1, \quad r \sim \left[\frac{2\tau}{3}(KN-M)\right]^{\frac{3}{2\tau}} \tag{33}$$

where $\alpha = \frac{2\tau-3}{2\tau}$, $\beta = \frac{3}{2\tau-3}$.

- *if $KN - M = O(1)$, then $l \to 1, r = \Theta(1)$, with the exact value determined by*

$$\begin{cases} KN-M+r-1 > (r-1)^{\frac{2\tau}{3}} H_{\frac{2\tau}{3}}(r-1), \\ \quad KN-M+r \leq r^{\frac{2\tau}{3}} H_{\frac{2\tau}{3}}(r). \end{cases} \tag{34}$$

Note that in the case of $\tau > {}^{3}/_{2}$, the asymptotic law for r is the same, $r = \Theta\left((KN-M)^{\frac{3}{2\tau}}\right)$ in both (32) and (33). Moreover, the approximation of (31) on $\hat{l}$ can be precisely carried out via (27), if we substitute K with $K - \lim {}^{M}/_{N}$.

TABLE III

(a) The Cases of $\tau < 1$, $\tau = 1$ and $1 < \tau < 3/2$.

M		M finite	$N\to\infty$, then $M\to\infty$	$M \overset{\lim}{\le} \frac{3-2\tau}{2\tau}KN$	$M \overset{\lim}{>} \frac{3-2\tau}{2\tau}KN$, and $M \overset{\lim}{<} KN$	$M \sim KN$	
						$KN-M=\omega(1)$	$KN-M=O(1)$
$\mathcal{M}_{\downarrow}$		empty	empty	almost empty	non-empty	non-empty	non-empty
$\hat{l}$		1	1	1	1	1	1
$\hat{r}$		$M+1$	$M+1$	$M-o(M)$	$\frac{3-2\tau}{2\tau}(KN-M)$	$\frac{3-2\tau}{2\tau}(KN-M)$	$\Theta(1)$ (34)
C	$\tau<1$	$\Theta(1)$	$\Theta\left(\sqrt{M}\right)$	$\Theta\left(\sqrt{M}\right)$	$\Theta\left(\sqrt{M}\right)$	$\Theta\left(\sqrt{M}\right)$	$\Theta\left(\sqrt{M}\right)$
	$\tau=1$	$\Theta(1)$	$\Theta\left(\frac{\sqrt{M}}{\log M}\right)$	$\Theta\left(\frac{\sqrt{M}}{\log M}\right)$	$\Theta\left(\frac{\sqrt{M}}{\log M}\right)$	$\Theta\left(\sqrt{M}\right)$	$\Theta\left(\sqrt{M}\right)$
	$1<\tau<\frac{3}{2}$	$\Theta(1)$	$\Theta\left(M^{\frac{3}{2}-\tau}\right)$	$\Theta\left(M^{\frac{3}{2}-\tau}\right)$	$\Theta\left(M^{\frac{3}{2}-\tau}\right)$	$\Theta\left(\frac{\sqrt{M}}{(KN-M)^{\tau-1}}\right)$	$\Theta\left(\sqrt{M}\right)$

(b) The Case of $\tau = 3/2$.

M	M finite	$N\to\infty$, then $M\to\infty$	$M\ln M \overset{\lim}{\le} KN$	$M\ln M \overset{\lim}{>} KN$ and $M \overset{\lim}{<} KN$	$M \sim KN$	
					$KN-M=\omega(1)$	$KN-M=O(1)$
$\mathcal{M}_{\downarrow}$	empty	empty	almost empty	non-empty	non-empty	non-empty
$\hat{l}$	1	1	1	1	1	1
$\hat{r}$	$M+1$	$M+1$	$M-o(M)$	$r\ln r \sim KN-M$	$r\ln r \sim KN-M$	$\Theta(1)$ (34)
C	$\Theta(1)$	$\Theta(\log^{\frac{3}{2}} M)$	$\Theta(\log^{\frac{3}{2}} M)$	$\Theta\left(\log^{\frac{3}{2}} r\right)$	$\Theta\left(\sqrt{\frac{M}{KN-M}}\log^{\frac{3}{2}} r\right)$	$\Theta\left(\sqrt{M}\right)$

(c) The Case of $\tau > 3/2$.

M	M finite	$N\to\infty$, then $M\to\infty$	$M \overset{\lim}{\le} hN^{\frac{3}{2\tau}}$ (see Th. 16)	$M \overset{\lim}{>} hN^{\frac{3}{2\tau}}$, and $M \overset{\lim}{\le} (K-\beta)N$	$M \overset{\lim}{>} (K-\beta)N$ and $M \overset{\lim}{<} KN$	$M \sim KN$	
						$KN-M=\omega(1)$	$KN-M=O(1)$
$\mathcal{M}_{\downarrow}$	empty	empty	almost empty	non-empty	non-empty	non-empty	non-empty
$\hat{l}$	$\Theta(1)$ (27)	$\Theta(1)$ (27)	$\Theta(1)$ (27)	$\cong\alpha\left[K+1-\lim \frac{M}{N}\right]$	1	1	1
$\hat{r}$	$M+1$	$M+1$	$M-o(M)$	$\sim\alpha\left[KN^{\frac{3}{2\tau}}-\frac{M}{N^{1-\frac{3}{2\tau}}}\right]$	$\sim\left[\frac{2\tau}{3}(KN-M)\right]^{\frac{3}{2\tau}}$	$\sim\left[\frac{2\tau}{3}(KN-M)\right]^{\frac{3}{2\tau}}$	$\Theta(1)$ (34)
C	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$	$\Theta(1)$	$\Theta\left(\frac{\sqrt{M}}{(KN-M)^{\frac{3(\tau-1)}{2\tau}}}\right)$	$\Theta\left(\sqrt{M}\right)$

Theorem 19 [Capacity for $M \overset{\lim}{<} KN$, $\mathcal{M}_{\downarrow} \neq \emptyset$]:

- *If* $\tau < 1$, $\quad C = \Theta\left(\sqrt{M}\right)$.
- *If* $\tau = 1$, $\quad C = \Theta\left(\frac{\sqrt{M}}{\log M}\right)$.
- *If* $1 < \tau < 3/2$, $C = \Theta\left(M^{\frac{3}{2}-\tau}\right)$.
- *If* $\tau = 3/2$, $\quad C = \Theta\left(\log^{\frac{3}{2}} r\right)$.
- *If* $\tau > 3/2$, $\quad C = \Theta\left(1\right)$.

Theorem 20 [Capacity for $M \sim KN$]:

- *If* $\tau \le 1$, $\quad C = \Theta\left(\sqrt{M}\right)$.
- *If* $1 < \tau < 3/2$, $C = \Theta\left(\frac{\sqrt{M}}{(KN-M)^{\tau-1}}\right)$.
- *If* $\tau = 3/2$, $\quad C = \Theta\left(\sqrt{\frac{M}{KN-M}}\log^{\frac{3}{2}} r\right)$.
- *If* $\tau > 3/2$, $\quad C = \Theta\left(\frac{\sqrt{M}}{(KN-M)^{\frac{3(\tau-1)}{2\tau}}}\right)$.

E. Validation for the WN Problem

Recapitulating on the asymptotic laws, C, $C^{\mathtt{CD}\star}$, and $C^{\mathtt{AN}\star}$ are of the same order (Theorem 9). To complete with the WN Problem, we consider the scaling of $A_{0,0}$; from (13),

$$A_{0,0} \le \frac{1}{K}\sum_{k=1}^{M} p_m^{\frac{2}{3}} \le \frac{1}{K} H_{\frac{2\tau}{3}}(M).$$

Thus, $A_{0,0}$ diverges only for $\tau \le 3/2$, as $M^{\tau-1}$ when $\tau < 3/2$, or $\log M$ when $\tau = 3/2$. As easily verified from Table III, $A_{0,0}$ always scales slower than C, therefore, Table III pertains to the scaling of the required rate for the worst link case, too.

F. Discussion on Asymptotic Laws

The main result of the asymptotic laws regards the minimum required link rate required to sustain a request rate of $\lambda = 1$ from each node. As a preliminary comment, the sustainable link rates are subject to the information theory and Shannon's capacity law. Thus, a rate C that scales to infinity should be interpreted rather as the inverse of the maximum sustainable request rate λ, e.g., the result of $C = \Theta(\sqrt{M})$ for $\lambda = 1$ is equivalent to $C = \Theta(1)$ for $\lambda = \Theta(1/\sqrt{M})$, as in [2].

The power law parameter τ sets two phase transition points, 1 and $3/2$, leading to distinct asymptotics: the higher τ,

the more uneven the popularity of files, and thus, the more advantage in caching (i.e. lower link rate C). As an example, on $\tau > 3/2$ and $M \leq (K-\epsilon)N$, with ϵ a small constant, $C = \Theta(1)$, or, in words, the wireless network is a sustainable. Nevertheless, such high a τ arises in particular scenarios (i.e., mobile applications), as discussed in Section II.

More common are the cases of low and intermediate values of τ (traffic in routers and proxies), which flatten the popularity distribution towards the uniform. Replication becomes less effective, ending up to the $\Theta\left(\sqrt{M}\right)$ law for $\tau < 1$, a synonym of the Gupta-Kumar law, if we associate the number of files M to the number of communicating pairs in [2]. When M scales slower than N, there is an improvement over [2], especially on $\tau \geq 1$, which, under the above prism, expresses the Gupta-Kumar law for the flows induced from the replication.

In an alternate perspective, the joint scaling of M and N can be considered as each new node enriching the network with its 'own' files. In the case of $M \sim \delta KN$ being a fraction of the total buffer capacity, the node has spare capacity to cache other files, too. The value of constant δ determines the size of the set $\mathcal{M}_{\downarrow}$ (Theorem 16), and, consequently, the scaling law of C: $\Theta\left(\sqrt{M}\right)$ for $\tau < 1$, $\Theta\left(M^{\frac{3}{2}-\tau}\right)$ for $1 < \tau < \frac{3}{2}$, or $\Theta(1)$ for $\tau > \frac{3}{2}$. Under this perspective, the improvement for $\tau \geq 1$ spotlights the advantage of replication: as M is of the same order with N, the flow model is a fair comparison to [2].

Last, when the ratio of M over KN approaches 1, hardly does it remain any capacity left for replication: rate C in the last two columns of the tables essentially match the Gupta-Kumar rate of $\Theta\left(\sqrt{M}\right)$.

VII. Conclusions & Future Work

In this work, we investigated on the joint delivery and replication problem of flat wireless networks with caching. After formulating the precise problem on square lattice networks, we gradually reduced it to a simple density-based problem; from its solution, we designed an efficient order-wise replication and delivery scheme and derived the scaling laws on the required link capacity. As our study reveals, the key parameters in network sustainability are the file popularity parameter τ and the relative scaling of the numbers of files vs. network nodes.

A future extension of this work will include a new dimension of scaling with node capacity K; that is, in an expanding network, not only new nodes and content are added, but existing nodes evolve with augmented buffer storage. In such a setup, replication is expected to be advantageous order-wise, even for low values of τ provided that cache capacity K scales sufficiently fast with the number of files M.

Appendix A

Proof of Lemma 3: Using Lemma 2, we use a single arbitrary shortest path for each n, m pair to a unique $w_{m,n}$. From the discussion on d_m, the set $\mathcal{W}_m$ has cardinality $W_m = Nd_m$, i.e., there exist Nd_m nodes in the network that serve the requests for file m. Let $\mathcal{Q}_{w,m}$ be the set of nodes served from node w for requests on m, and $Q_{w,m}$ be its cardinality. Clearly,

$$\sum_{w\in\mathcal{W}_m} Q_{w,m} = N. \tag{35}$$

It is not hard to see that the best arrangement for a node w that serves a cluster of $Q_{w,m}$ nodes on requests for data m, is when these nodes lie in a square rhombus centered at w: this minimizes the total hop-count $\sum_{n\in\mathcal{Q}_{w,m}} h(n,m)$. Thus, the 1st node ($w$ itself) has a hop count $h=0$, the next 4 nodes have $h=1$, the next 8 nodes have $h=2$, etc. as illustrated in Fig. 4. Therefore, taking into account the nodes of the last incomplete ring at hop count equal to $\check{\rho}+1$, it is

$$\begin{aligned}\sum_{n\in\mathcal{Q}_{w,m}} h(n,m) \geq\ & \overbrace{1\times 0 + 4\times 1 + 8\times 2 + \cdots + 4\check{\rho}\cdot\check{\rho}}^{\text{Hops for the nodes of the rhombus at hops } \leq\check{\rho}} \\ & + \overbrace{\left[Q_{w_m} - 1 - 2(\check{\rho}+1)\check{\rho}\right]}^{\text{Number of nodes at } \check{\rho}+1}\overbrace{(\check{\rho}+1)}^{\times(\check{\rho}+1)} \\ =\ & 2\check{\rho}(\check{\rho}+1)\frac{2\check{\rho}+1}{3} + [Q_{w,m}-1-2(\check{\rho}+1)\check{\rho}](\check{\rho}+1),\end{aligned}$$

where $\check{\rho}$ is the radius of the rhombus. The radius can be computed as the integer part $\check{\rho} = \lfloor\rho\rfloor$, where ρ satisfies

$$Q_{w,m} = 2(\rho+1)\rho + 1.$$

Indeed, the RHS expresses the number of elements in a rhombus of radius ρ, when ρ is an integer. Solving for ρ,

$$\rho = \frac{-1+\sqrt{2Q_{w,m}-1}}{2}. \tag{36}$$

Observe that

$$2\rho(\rho+1)\frac{2\rho+1}{3} \leq 2\check{\rho}(\check{\rho}+1)\frac{2\check{\rho}+1}{3} + [Q_{w,m}-1-2(\check{\rho}+1)\check{\rho}](\check{\rho}+1). \tag{37}$$

Indeed, the above is an equality when ρ is integer, as the second term of the RHS vanishes. Moreover, it is easy to see that the LHS is a convex and increasing function of $Q_{w,m}$, (i.e., if we substitute $\rho = \rho(Q_{w,m})$ from (36)), whereas the RHS is piecewise linear on $Q_{w,m}$ (the pieces' endpoints correspond to integer values of ρ). Comparing the two, the LHS cannot exceed the RHS, as illustrated in Fig. 5. Thus,

$$\sum_{n\in\mathcal{Q}_{w,m}} h(n,m) \geq 2\rho(\rho+1)\frac{2\rho+1}{3} = \frac{\sqrt{2}}{3}\left(Q_{w,m}-1\right)\sqrt{Q_{w,m}-\frac{1}{2}},$$

with the equality being true when ρ is integer. Moreover, since $f(x) = (x-1)\sqrt{x-\frac{1}{2}} - x(\sqrt{x}-1) \geq 0$ for $x \geq 1$ (as $f(1)=0$, $f'(1)>0$ and $f''(x)\geq 0$ for $x\geq 0$), it is

$$\sum_{n\in\mathcal{Q}_{w,m}} h(n,m) \geq \frac{\sqrt{2}}{3}Q_{w,m}\left[Q_{w,m}^{\frac{1}{2}} - 1\right]. \tag{38}$$

Then, the total number of hops for each m is expressed as

$$\begin{aligned}\sum_{n\in\mathcal{N}} h(n,m) &= \sum_{w\in\mathcal{W}_m}\sum_{n\in\mathcal{Q}_{w,m}} h(n,m) \geq \sum_{w\in\mathcal{W}_m}\frac{\sqrt{2}}{3}Q_{w,m}\left[Q_{w,m}^{\frac{1}{2}}-1\right] \\ &\geq \frac{\sqrt{2}}{3}W_m\left[\sum_{w\in\mathcal{W}_m} Q_{w,m}W_m^{-1}\right]\left[\sqrt{\sum_{w\in\mathcal{W}_m} Q_{w,m}W_m^{-1}}-1\right] \\ &\overset{(35)}{=} \frac{\sqrt{2}}{3}N\left[d_m^{-\frac{1}{2}}-1\right]\end{aligned}$$

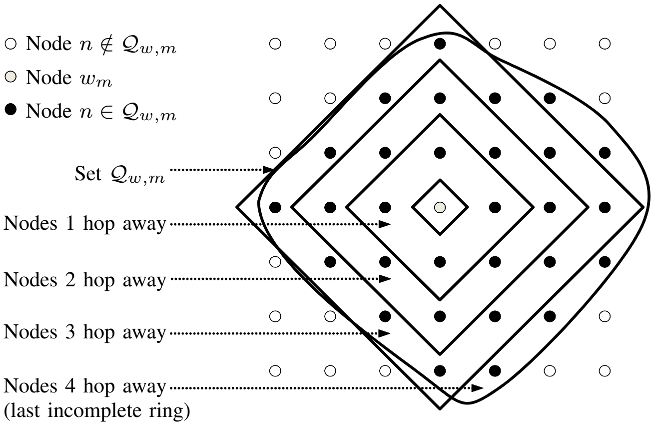

Fig. 4. A cluster of peers served by node w_m for data m of $\check{\rho} = 3$.

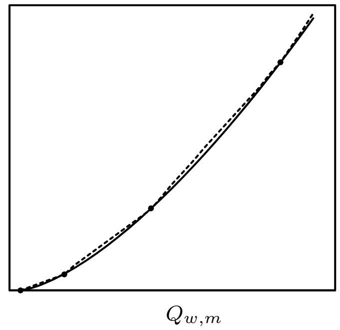

Fig. 5. The RHS of (37)) (dashed line) and the LHS of (37) versus the cluster size $Q_{w,m}$: knots indicate the points of $\rho(Q_{w,m}) = 1, 2, \ldots$, where (37) is an equality.

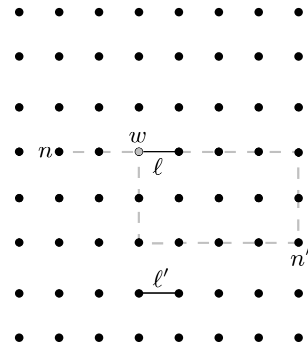

Fig. 6. The square cluster $\mathcal{Q}_w$ of the nodes served by node w, with the I-shaped and L-shaped routes for nodes n and n', respectively.

where the last inequality is Jensen's inequality applied on the convex function $g(x) = x\left(\sqrt{x} - 1\right)$ for the average node count per w, $\sum_w Q_{w,m} W_m^{-1}$.

In total,

$$\operatorname*{ave}_{\ell} C_\ell = \frac{1}{2N} \sum_{m\in\mathcal{M}} \sum_{n\in\mathcal{N}} h(n,m) p_m \geq \frac{\sqrt{2}}{6} \sum_{m\in\mathcal{M}} \left(d_m^{-\frac{1}{2}} - 1\right) p_m.$$

As this inequality is true for the optimal values of the AT problem, the result follows. ■

Proof of Lemma 5: Consider density vectors $[d_m]$ and $[d'_m]$ both satisfying the constraints, and their convex combination $d''_m = \lambda d_m + (1-\lambda) d'_m$ for $\lambda \in (0,1)$. Function $\frac{1}{\sqrt{x}}$ is strictly convex, hence $C^{\text{CD}}([d''_m]) < \lambda C^{\text{CD}}([d_m]) + (1-\lambda) C^{\text{CD}}([d'_m])$, i.e., the objective $C^{\text{CD}}([d_m])$ is strictly convex. Moreover, the constraints $N^{-1} \leq d_m \leq 1$ are convex, and the constraint $\sum_{m\in\mathcal{M}} d_m \leq K$ is affine. Thus, the optimization is strictly convex [20, §4.2.1], and has a unique optimal solution. ■

Proof of Lemma 6: Let us first prove that $\mathcal{M}_{\uparrow} = \{1, 2, \ldots, l-1\}$. Suppose that in the optimal solution $d_m^{\text{CD}\star}$, $\mathcal{M}_{\uparrow}$ contains m_2 but not m_1, with $m_1 < m_2$. Due to the ordering, it is $p_{m_1} \geq p_{m_2}$. We then construct a new solution d'_m by exchanging the densities of m_1 and m_2, i.e. $d'_{m_2} = d_{m_1}^{\text{CD}\star}$ and $d'_{m_1} = d_{m_2}^{\text{CD}\star}$, and maintaining the densities of any other m: $d'_m = d_m^{\text{CD}\star}$ for $m \in \mathcal{M} \setminus \{m_1, m_2\}$.

The link rate of the two solutions differ by

$$\begin{aligned} C^{\text{CD}\star} - C' &= \frac{\sqrt{2}}{6}\left[\left(\frac{p_{m_1}}{\sqrt{d_{m_1}^{\text{CD}\star}}} - \frac{p_{m_2}}{\sqrt{d'_{m_2}}}\right) + (p_{m_2} - p_{m_1})\right] \\ &= \frac{\sqrt{2}}{6}(p_{m_1} - p_{m_2})\left(\frac{1}{\sqrt{d_{m_1}^{\text{CD}\star}}} - 1\right). \end{aligned}$$

If $p_{m_1} = p_{m_2}$, d'_m has the same link rate C^{CD} as the $d_m^{\text{CD}\star}$, therefore it is optimal, too. This means that we can replace $d_m^{\text{CD}\star}$ with d'_m to construct a second optimal solution, which is a contradiction as the optimal solution is unique (Lemma 5).

Consider now the case of $p_{m_1} > p_{m_2}$. As $m_1 \notin \mathcal{M}_{\uparrow}$, it is $d_{m_1}^{\text{CD}\star} = d'_{m_2} < 1$. Clearly, from the above equation, $p_{m_1} > p_{m_2}$ implies that $C^{\text{CD}\star} - C'^{\text{CD}} > 0$. Thus, $d_{m_1}^{\text{CD}\star}$ is not optimal, which is a contradiction to the hypothesis about m_1 and m_2.

Similarly, for the $\mathcal{M}_{\downarrow}$ part, a solution with $m_3 \in \mathcal{M}_{\downarrow}$, $m_4 \notin \mathcal{M}_{\downarrow}$ and $m_3 < m_4$ leads to a contradiction. ■

Proof of Theorem 7: The first part has already been shown. For the second part, note that $d_m^{\text{CD}\star} < 4 d_m^\circ$. Thus,

$$\begin{aligned} 2C^{\text{CD}\star} &= \frac{\sqrt{2}}{3}\sum_{m\in\mathcal{M}}\left[\frac{1}{\sqrt{d_m^{\text{CD}\star}}} - 1\right] p_m = \frac{\sqrt{2}}{6}\sum_{m\in\mathcal{M}}\left[\sqrt{\frac{4}{d_m^{\text{CD}\star}}} - 2\right] p_m \\ &> \frac{\sqrt{2}}{6}\sum_{m\in\mathcal{M}}\left[\frac{1}{\sqrt{d_m^\circ}} - 2\right] p_m = C^{\text{CD}\circ} - \frac{\sqrt{2}}{6}. \end{aligned}$$ ■

Proof of Lemma 8: Observe that on each iteration of k, Algorithm 1 adds data to a cache with the least items. Thus, for cache n to end up with $K+1$ or more items, all other $n' \neq n$ should have at least K items, i.e., $\sum_{n\in\mathcal{N}} |\mathcal{B}_n| \geq NK + 1$.

This is a contraddiction: each item m appears in $N d_m^\circ = 4^{\nu - \nu_m}$ caches, and from d_m° construction, it is

$$\sum_{n\in\mathcal{N}} |\mathcal{B}_n| = N \sum_{m\in\mathcal{M}} d_m^\circ \overset{(12)}{\leq} N \sum_{m\in\mathcal{M}} d_m^{\text{CD}\star} \leq NK.$$ ■

Proof of Theorem 9: Given the shortest path routing and the replication pattern, we can compute the expected number of hop counts from a random client to the node serving its request. Indeed, consider that a node in $\mathcal{W}_m$ that stores m in its cache, and let us compute the summation of hops to the client nodes in $\mathcal{Q}_{w,m}$ it serves. $\mathcal{Q}_{w,m}$ is a set of nodes laying on a square of size $2^{\nu_m^\circ} \times 2^{\nu_m^\circ}$, with w_m at its center (Fig. 6). From symmetry, we can perform the summation by counting twice the required hops along one axis. For $d_m^\circ < 1$,

$$\begin{aligned} \sum_{n\in\mathcal{Q}_{w,m}} h(n,m) &= 2\times\overbrace{[1+\ldots+2^{\nu_m^\circ - 1} + 1 + \ldots + (2^{\nu_m^\circ - 1} - 1)]}^{\text{Movements on the vertical direction}}\overbrace{2^{\nu_m^\circ}}^{\text{columns}} \\ &= 2^{3\nu_m^\circ - 1}. \end{aligned}$$

Given the $N d_m^\circ$ nodes of $\mathcal{W}_m$, the average load across all links is computed by summing across m from (9):

$$\begin{aligned} C^{\text{AN}\circ} &= \frac{1}{2N}\sum_{m\in M}\sum_{w\in\mathcal{W}_m} 2^{3\nu_m^\circ - 1}\mathbb{1}_{\{d_m^\circ < 1\}} p_m \\ &= \frac{1}{2N}\sum_{m\in M} N d_m^\circ 2^{3\nu_m^\circ - 1}\mathbb{1}_{\{d_m^\circ < 1\}} p_m \\ &\leq \frac{1}{4} + \frac{1}{4}\sum_{m\in M}\left[\frac{1}{\sqrt{d_m^\circ}} - 1\right] p_m = \frac{1}{4} + \frac{3}{4}\sqrt{2}C^{\text{CD}\circ}. \end{aligned}$$

Given that (Theorem 4) $C^{\texttt{CD}\star} \leq C^{\texttt{AN}\star} \leq C^{\texttt{AN}\circ}$, and (Theorem 7) $C^{\texttt{CD}\circ} \leq 2C^{\texttt{CD}\star} + \sqrt{2}/6$, it follows that

$$C^{\texttt{AN}\star} \leq C^{\texttt{AN}\circ} \leq \frac{1}{2} + \frac{3}{2}\sqrt{2}C^{\texttt{CD}\star} \leq \frac{1}{2} + \frac{3}{2}\sqrt{2}C^{\texttt{AN}\star}. \qquad \blacksquare$$

Proof of Theorem 10: Consider a link ℓ of the network and a data m, and let us calculate the load $C_{\ell,m}$ incurred on ℓ for requests on data m. Traffic regarding m flows through link ℓ only if ℓ links nodes of the same cluster $\mathcal{Q}_{w,m}$. Next Lemma computes a bound on the load $C_{\ell,m}$. Note that Algorithm 1 constructs a square-shaped cluster $\mathcal{Q}_{w,m}$ of size $2^{\nu_m^\circ} \times 2^{\nu_m^\circ}$.

LEMMA 21 [LINK LOAD PER DATA]:

- *If link ℓ is adjacent to nodes belonging to different clusters $\mathcal{Q}_{w_m}$ or $\nu_m^\circ = 0$, then $C_{\ell,m} = 0$.*
- *If ℓ is in the same row or column with w_m, then $C_{\ell,m} \leq 2^{\nu_m^\circ - 1}\left(2^{\nu_m^\circ - 2} + \frac{1}{2}\right) p_m$.*
- *If ℓ is not in the same row or column with w_m, then $C_{\ell,m} \leq 2^{\nu_m^\circ - 2} p_m$.*

Proof: Fig. 6 depicts links ℓ and ℓ', which constitute the worst case (w.r.t. carried traffic) for row links in the same and different rows to w_m (column links can be treated similarly).

Clearly, for link ℓ, factor $2^{\nu_m^\circ - 1}$ counts the number of columns that lie from ℓ and outwards to the boundary of $\mathcal{Q}_{w,m}$, $2^{\nu_m^\circ - 2} + \frac{1}{2}$ is the number of nodes in each of these columns, with $1/2$ accounting for two-path routing for the nodes in different rows than ℓ, and p_m is the popularity of m.

Similarly, for link ℓ', $2^{\nu_m^\circ - 2}$ is the number $2^{\nu_m^\circ - 1}$ of nodes that are served through link ℓ' multiplied by $1/2$ due to two-path routing, and p_m is the file popularity. $\blacksquare$

Using the above, we bound the load of an arbitrary link ℓ, considering the case of l being a row link (column links can be treated similarly), located at row y_ℓ (y_ℓ takes values from 1 to 2^ν) summing over all data m that have $\nu_m^\circ > 0$:

$$\begin{aligned}
C_\ell = \sum_{m\in\mathcal{M}} C_{\ell,m} \leq & \sum_{m\in\mathcal{M}\setminus\mathcal{M}_0} p_m \Bigg[2^{\nu_m^\circ - 1}\mathbb{1}_{\{\nexists w_m \text{in the row of } \ell\}} \\
& + 2^{\nu_m^\circ - 1}\left(2^{\nu_m^\circ - 1} + \frac{1}{2}\right)\mathbb{1}_{\{\exists w_m \text{in the row of } \ell\}}\Bigg] \\
\leq & \left[\sum_{m\in\mathcal{M}\setminus\mathcal{M}_0} 2^{\nu_m^\circ - 1} p_m\right] + \sum_{m\in\mathcal{M}\setminus\mathcal{M}_0} 2^{2\nu_m^\circ - 2} p_m \mathbb{1}_{\{\exists w_m \text{in the row of } \ell\}}
\end{aligned}$$

As $1/\sqrt{d_m^\circ} = 2^{\nu_m^\circ}$, the first summation is less or equal to $2C^{\texttt{AN}\circ} + 2$. As for the second term, Lemma 23 that follows, bounds it by $2C^{\texttt{AN}\circ} + 2 + A_{0,0}/4$. Using Theorem 9 as well,

$$C^{\texttt{WN}\circ} = \max_l C_l \leq 4C^{\texttt{AN}\circ} + 2 + \frac{A_{0,0}}{4} \leq 2 + \frac{A_{0,0}}{4} + 6\sqrt{2}C^{\texttt{AN}\star}.$$

Given, moreover, that $C^{\texttt{AN}\star} \leq C^{\texttt{WN}\star}$, the proof is complete. $\blacksquare$

Before establishing the bound of Lemma 23, we first make the following observation on the workings of Algorithm 1:

LEMMA 22: *For every row y, and at every step of Algorithm 1's run, each replicated file that lies in the buffers of the nodes at row y_ℓ can be uniquely matched to another previously added replicated file in the buffers of the nodes at row y, excluding from the matchings the files of $\mathcal{M}_0$ and the replicas of the first file from the set $\mathcal{M} \setminus \mathcal{M}_0$ added to row y_ℓ.*

Proof: Observe that the Algorithm operates on the $2^k \times 2^k$ submatrix filling in data to the diagonals from set $\mathcal{M}_k$. When $\mathcal{M}_k$ gets exhausted, the algorithm moves to $\mathcal{M}_{k+1}$ operating on the $2^{k+1} \times 2^{k+1}$ formed by repeating the $2^k \times 2^k$ submatrix. Thus, if we exclude the files of $\mathcal{M}_0$, duplicated to all nodes, and the first time we visit row y_ℓ to add a file[2] (this will regard node (y_ℓ, y_ℓ) of the first diagonal), we know for sure that by the time we revisit row y_ℓ (at the node of the second diagonal), we will have already visited all other rows. Thus, the file added once to the first $2^k \times 2^k$ submatrix can be matched uniquely to another element previously put in the buffer of row y. Similarly, on the third and so forth visit of row y_ℓ to add a new file to the first $2^k \times 2^k$ submatrix, we will have already have revisited row y to add another file in the the first $2^k \times 2^k$ submatrix, and thus we can match these two files.

The above is true for the whole $2^k \times 2^k$ network, too. When $\mathcal{M}_k$ gets exhausted, we replicate the $2^k \times 2^k$ submatrix to the $2^{k+1} \times 2^{k+1}$ submatrix (step 4); in this process, the matchings in the $2^k \times 2^k$ submatrix between the replicas of files in the buffers of row y_ℓ to the ones in row y_l are replicated to the 3 new $2^k \times 2^k$ submatrices of the large $2^{k+1} \times 2^{k+1}$ submatrix. Thus, we established the matching of each replicated file in row y_ℓ excluding the files of $\mathcal{M}_0$ and the replicas of the first file from $\mathcal{M} \setminus \mathcal{M}_0$ to the replicated files in row y. $\blacksquare$

LEMMA 23: *For any row $y_0 = 1, 2, \ldots, 2^\nu$,*

$$4 \operatorname*{ave}_{\ell} C_l + \frac{A_{0,0}}{8} \geq \sum_{m\in\mathcal{M}\setminus\mathcal{M}_0} 2^{2\nu_m^\circ - 3} p_m \mathbb{1}_{\{m \textit{ exists in row } y\}}.$$

Proof: Continuing Lemma 22, let $\hat{m}_j$ be the file added in row y_ℓ with j denoting their order, excluding set $\mathcal{M}_0$. Each one appears $4^{\nu - \nu_{\hat{m}_j}^\circ}$ times in the network, and $2^{\nu - \nu_{\hat{m}_j}^\circ}$ times in row y_ℓ. Let us index these $2^{\nu - \nu_{\hat{m}_j}^\circ}$ replicas of $\hat{m}_j$ in row y_ℓ for $j \geq 2$, by $r = 1, 2, \ldots, 2^{\nu - \nu_{\hat{m}_j}^\circ}$. In the matching of Lemma 22, remember that we excluded the replicas of the first data added out of $\mathcal{M} \setminus \mathcal{M}_0$, that is, the replicas of $\hat{m}_1$.

Let, moreover, $\hat{m}_{j,y,r}$ be the matching of the r replica of $\hat{m}_j$ at row y_ℓ to the (replicated) data at row y. As already discussed, $\hat{m}_{j,y,r}$ has been added to the network before $\hat{m}_j$. This observation creates the following cases:

Case 1: Both $\hat{m}_j$ and $\hat{m}_{j,y,r}$ belong to set $\mathcal{M}_{\updownarrow}$. Then,

$$1 = p_{\hat{m}_j}^{-1} p_{\hat{m}_j} = \left[A_{M_\uparrow, M_\downarrow} d_{\hat{m}_j}^{\texttt{CD}\star}\right]^{-\frac{3}{2}} p_{\hat{m}_j} \overset{(12)}{\geq} \left[4A_{M_\uparrow, M_\downarrow} d_{\hat{m}_j}^\circ\right]^{-\frac{3}{2}} p_{\hat{m}_j},$$

and similarly, $1 = p_{\hat{m}_{j,y,r}}^{-1} p_{\hat{m}_{j,y,r}} =$

$$= \left[A_{M_\uparrow, M_\downarrow} d_{\hat{m}_{j,y,r}}^{\texttt{CD}\star}\right]^{-\frac{3}{2}} p_{\hat{m}_{j,y,r}} \leq \left[A_{M_\uparrow, M_\downarrow} d_{\hat{m}_{j,y,r}}^\circ\right]^{-\frac{3}{2}} p_{\hat{m}_{j,y,r}}.$$

Combining these two inequalities, and using $d_m^\circ = 4^{-\nu_m^\circ}$,

$$2^{3\nu_{\hat{m}_j}^\circ - 3} p_{\hat{m}_j} \leq 2^{3\nu_{\hat{m}_{j,y,r}}^\circ} p_{\hat{m}_{j,y,r}}. \tag{39}$$

Case 2: $\hat{m}_j \in \mathcal{M}_\downarrow$, $\hat{m}_{j,y,r} \in \mathcal{M}_\updownarrow$. Clearly, $\nu_{\hat{m}_j}^\circ = \nu$, whereas $\nu_{\hat{m}_{j,y,r}}^\circ \in \{1, 2, \ldots, \nu\}$. The fact that $\hat{m}_{j,y,r} \in \mathcal{M}_\updownarrow$ imposes a constraint on the probabilities $p_{\hat{m}_j}$ vs. $p_{\hat{m}_{j,y,r}}$ of (11b):

$$p_{\hat{m}_{j,y,r}} \geq 8^{\nu - \nu_{\hat{m}_{j,y,r}}^\circ} p_{\hat{m}_j} = 2^{3\nu - 3\nu_{\hat{m}_{j,y,r}}^\circ} p_{\hat{m}_j},$$

[2] To be precise, this will happen when we first visit row $y_\ell \bmod 2^k$—the contents of this row will get repeated to row y_ℓ when k becomes $2^k > y_\ell$.

i.e., if $\nu^{\circ}_{\hat{m}_{j,y,r}} = \nu$, or equivalently $N^{-1} \leq d^{\texttt{CD}\star}_{\hat{m}_{j,y,r}} < 4N^{-1}$, the above reads as $p_{\hat{m}_{j,y,r}} > p_{\hat{m}_j}$ as expected. However, if, for example $\nu^{\circ}_{\hat{m}_{j,y,r}} = \nu - 1$, or equivalently $4N^{-1} \leq d^{\texttt{CD}\star}_{\hat{m}_{j,y,r}} < 16N^{-1}$, it has to be $p_{\hat{m}_{j,y,r}} > 8p_{\hat{m}_j}$ so that (11b) is valid. Hence, (39) holds in this case, too.

Case 3: Both $\hat{m}_j$ and $\hat{m}_{j,y,r}$ belong to $\mathcal{M}_{\downarrow}$. Then, $\nu^{\circ}_{\hat{m}_{j,y,r}} = \nu^{\circ}_{\hat{m}_j}$. Then, given that $p_{\hat{m}_{j,y,r}} \geq p_{\hat{m}_j}$, (39) is valid, too.

Summing (39) for all y, $j \geq 2$ and $r = 1, 2, \ldots, 2^{\nu - \nu^{\circ}_{\hat{m}_j}}$,

$$
\begin{aligned}
2^{\nu} \sum_{j\geq 2} \sum_{r \geq 1} 2^{3\nu^{\circ}_{\hat{m}_j} - 3} p_{\hat{m}_j} &\leq \sum_{y} \sum_{j \geq 2} \sum_{r \geq 1} 2^{3\nu^{\circ}_{\hat{m}_{j,y,r}}} p_{\hat{m}_{j,y,r}} \Leftrightarrow \\
4^{\nu} \sum_{j \geq 2} 2^{2\nu^{\circ}_{\hat{m}_j} - 3} p_{\hat{m}_j} &\leq \sum_{y} \sum_{j \geq 2} \sum_{r \geq 1} 2^{3\nu^{\circ}_{\hat{m}_{j,y,r}}} p_{\hat{m}_{j,y,r}}, \qquad (40)
\end{aligned}
$$

where in the last step we used the number $2^{\nu - \nu^{\circ}_{\hat{m}_j}}$ of replicas of each $\hat{m}_j$ on row y_ℓ on the LHS to eliminate the inner sum.

We do the same for the RHS: the triple summation encompasses almost all the replicas placed in the network. The 'almost' regards the replicas added in the network after the addition of the last file $\hat{m}_j$ in row y_ℓ which do not participate to the matching. Thus, we can increase the RHS to encompass all replicas in the network. As each file $m \in \mathcal{M} \setminus \mathcal{M}_0$ has $4^{\nu - \nu^{\circ}_m}$ replicas in the network, the RHS is upper bounded by

$$
\sum_{m \in \mathcal{M} \setminus \mathcal{M}_0} 4^{\nu - \nu^{\circ}_m} 2^{3\nu^{\circ}_m} p_m \overset{2^{2\nu} = N}{=} N \sum_{m \in \mathcal{M} \setminus \mathcal{M}_0} 2^{-\nu^{\circ}_m} p_m \leq N \big(C^{\texttt{AN}\circ} + 1 \big) .
$$

Using the above bounds on the RHS with the LHS of (40),

$$
\begin{aligned}
4^{\nu} \sum_{j \geq 2} 2^{2\nu^{\circ}_{\hat{m}_j} - 3} p_{\hat{m}_j} &\leq N \left(C^{\texttt{AN}\circ} + 1 \right) \Leftrightarrow \\
\sum_{j} 2^{3\nu^{\circ}_{\hat{m}_j} - 2} p_{\hat{m}_j} &\leq 2C^{\texttt{AN}\circ} + 2 + 2^{2\nu^{\circ}_{\hat{m}_1} - 2} p_{\hat{m}_1}
\end{aligned}
$$

Last, consider the term of $\hat{m}_1 \in \mathcal{M} \setminus \mathcal{M}_0$ in the LHS:

$$
2^{2\nu^{\circ}_{\hat{m}_1}} p_{\hat{m}_1} = \frac{p_{\hat{m}_1}}{d^{\circ}_{\hat{m}_1}} \leq \frac{p_{\hat{m}_1}}{d^{\texttt{CD}\star}_{\hat{m}_1}} \leq \frac{A_{M_{\uparrow}, M_{\downarrow}}}{p^{\frac{2}{3}}_{\hat{m}_1}} p_{\hat{m}_1} = A_{M_{\uparrow}, M_{\downarrow}} p^{\frac{1}{3}}_{\hat{m}_1} \leq A_{0,0},
$$

where the first inequality is equality or strict inequality depending on whether $\hat{m}_1 \in \mathcal{M}_{\uparrow}$ or $\hat{m}_1 \in \mathcal{M}_{\downarrow}$, respectively.

Combining the last two inequalities, the result follows. ■

Appendix B

Proof of Lemma 11: If we assume that in the limit $l \overset{\lim}{>} 1$, then we have two cases for r: $r \to \infty$, or $r = O(1)$.

If $r \to \infty$, $\tau \leq 3/2$ leads to $H_{\frac{2\tau}{3}}(r-1)$ diverging to infinity in (22). However, the rest of the terms in (22) are bounded (as $l \leq K$). Therefore, (22) is a contradiction. Thus, it has to be either $\mathcal{M}_{\uparrow} = \emptyset$, or $l = 1$. As $r \to \infty$ and $l \leq K+1$, it cannot be $\mathcal{M}_{\uparrow} = \emptyset$. Therefore, if $r \to \infty$, it is $l = 1$ (i.e., $d_l \ncong 1$).

If $r = O(1)$, (26) contradicts with $l \overset{\lim}{>} 1$. Thus, $l \to 1$. ■

Proof of Lemma 12: Follows from the summation definition of (17): observe that $d_m > N^{-1}$ and $\sum_{m \in \mathcal{M}_{\uparrow}} p_m \leq 1$. ■

Proof of Lemma 13: For $\tau < 1$, $C_{\downarrow} = \Theta\left(\sqrt{N}\right)$ follows (18) and the fact that $H_\tau(M) - H_\tau(r) \overset{(15)}{=} \Theta\left(H_\tau(M)\right)$. The latter comes from $H_\tau(M)$ diverging and $r \overset{\lim}{<} M$.

For $\tau > 1$, it is $C_{\downarrow} \overset{(15)}{=} \Theta\Big(\sqrt{N}\left[r^{1-\tau} - M^{1-\tau}\right]\Big)$.

If $r \overset{\lim}{<} M$, too, then $M^{1-\tau} \overset{\lim}{<} r^{1-\tau}$, hence $C_{\downarrow} = \Theta\Big(\frac{\sqrt{N}}{r^{\tau-1}}\Big)$. ■

Proof of Theorem 15: **Case** $\tau \leq \frac{3}{2}$: From Lemma 11, $l \to 1$. **Case** $\tau > \frac{3}{2}$: Examining what happens in (20)-(21) in the limit, we observe that $r \to \infty$, hence $H_{\frac{2\tau}{3}}(r-1) \to H_{\frac{2\tau}{3}}$. Assuming a limit $l \to \hat{l}_{\{\tau > \frac{3}{2}, \mathcal{M} \approx \emptyset\}}$, (20)-(21) lead to (27); the latter, as in Lemma 5, have a unique solution. ■

Proof of Theorem 16: By the definition of $\mathcal{M}_{\downarrow}$ almost empty, it is $M - r + 1 = o(M)$, and given the constraint of (1), it is $M = O(N)$, thus $M - r + 1 = o(N) = o((K - l + 1)N)$, as $K - l + 1 \geq 1$ in all cases from Theorem 15.

From the last element of $\mathcal{M}_{\uparrow}$, we have that $d_{r-1} > N^{-1}$. Substituting d_{r-1} in the latter from (11b), and taking the limit

$$
(K - l + 1)N > (r-1)^{\frac{2\tau}{3}} \Big[H_{\frac{2\tau}{3}}(r-1) - H_{\frac{2\tau}{3}}(l-1) \Big] , \qquad (41)
$$

where we used $M - r + 1 = o((K - l + 1)N)$ to eliminate the respective term from the LHS. Next, we use (15) to approximate the Riemann terms and substitute l from Theorem 15:

Case $0 < \tau < 3/2$: $l \to 1$, thus (41) becomes

$$
KN \overset{\lim}{\geq} (r-1)^{\frac{2\tau}{3}} \frac{(r-1)^{1-\frac{2\tau}{3}} - 1}{1 - \frac{2\tau}{3}} = \frac{r - 1 - (r-1)^{\frac{2\tau}{3}}}{1 - \frac{2\tau}{3}} .
$$

As $2\tau/3 < 1$, it is $(r-1)^{\frac{2\tau}{3}} = o\,(r-1)$. Hence, the above is equivalent in the limit to $(r-1) \overset{\lim}{\leq} K\left(1 - \frac{2\tau}{3}\right) N$, or, as $r = \Theta(M)$, $M \overset{\lim}{\leq} K\left(1 - \frac{2\tau}{3}\right) N$.

Case $\tau = 3/2$: $l \to 1$, thus $KN \overset{\lim}{\geq} (r-1)\left[\ln(r-1) - \ln l\right]$. Using that $\ln l = o(\ln(r-1))$, we get $(r-1)\ln(r-1) \overset{\lim}{\leq} KN$; as $r = \Theta(M)$, the condition becomes $M \ln M \overset{\lim}{\leq} KN$.

Case $\tau > 3/2$:

$$
(K - \hat{l} + 1)N \overset{\lim}{\geq} (r-1)^{\frac{2\tau}{3}} \frac{\hat{l}^{1-\frac{2\tau}{3}} - (r-1)^{1-\frac{2\tau}{3}}}{\frac{2\tau}{3} - 1} = \frac{\hat{l}^{1-\frac{2\tau}{3}} (r-1)^{\frac{2\tau}{3}} - (r-1)}{\frac{2\tau}{3} - 1} .
$$

As $2\tau/3 > 1$, it follows that $(r-1) = o\Big((r-1)^{\frac{2\tau}{3}}\Big)$, and the above becomes $(r-1) \overset{\lim}{\leq} \left[\frac{\left(K - \hat{l} + 1\right)\left(\frac{2\tau}{3} - 1\right)}{\hat{l}^{1-\frac{2\tau}{3}}} N \right]^{\frac{3}{2\tau}}$. Substituting $r - 1$ with M, the condition follows.

Last, observe that in the above derivations, if we started with $d_M > N^{-1}$, we would find the conditions for $\mathcal{M}_{\downarrow}$ being strictly empty, i.e. $r = M + 1$. As easily seen, this is true if the conditions are satisfied with strict inequality. ■

Proof of Theorem 17: To find C, we compute $C_{\uparrow}$ and $C_{\downarrow}$ from (17)-(18), and show that in all cases $C_{\downarrow} = O\Big(C_{\uparrow}\Big)$. Thus, $C = \Theta\Big(C_{\uparrow}\Big)$. In assessing $C_{\downarrow}$, $r \sim M$ helps in deriving that

$$
H_\tau(M) - H_\tau(r-1) = \sum_{j=r}^{M} j^{-\tau} = \Theta\big(M^{-\tau}(M - r)\big) .
$$

Theorem 15 and $M - r = o(M)$ result in $K_{\uparrow} = \Theta(1)$ for all τ.

Case $\tau < 1$: Regarding $C_{\uparrow}$, $H_{\frac{2\tau}{3}}(r+1)$ and $H_\tau(M)$ diverge, while $H_{\frac{2\tau}{3}}(l-1)$ is bounded (as $l \leq K+1$). Thus,

$$
C_{\uparrow} = \Theta\left(\frac{\left[M^{1-\frac{2\tau}{3}} - 1\right]^{\frac{3}{2}}}{M^{1-\tau} - 1} \right) = \Theta\Big(\sqrt{M}\Big) .
$$

If the condition of Theorem 16 is a strict inequality, $C_{\downarrow} = 0$.

Otherwise, it is an equality, with $M=\Theta(N)$, and thus,

$$C_{\downarrow}=\sqrt{N}\frac{\sum_{j=r}^{M}j^{-\tau}}{H_\tau(M)}\overset{(15)}{=}\Theta\Big(\sqrt{N}\frac{M^{-\tau}(M-r)}{M^{1-\tau}}\Big)\overset{M=\Theta(N)}{=}o\Big(\sqrt{M}\Big).$$

Case $\tau=1$: In $C_{\updownarrow}$, $H_{\frac{2\tau}{3}}(M)$ and $H_\tau(M)$ diverge, while $H_{\frac{2\tau}{3}}(l-1)$ is bounded. Thus, $C_{\updownarrow}=\Theta\left(\frac{\left(M^{\frac{1}{3}}-1\right)^{\frac{3}{2}}}{\log M}\right)=\Theta\Big(\frac{\sqrt{M}}{\log M}\Big)$.
As before, if the condition of Theorem 16 is a strict inequality, $C_{\downarrow}=0$. Otherwise, it is an equality, with $M=\Theta(N)$, thus,

$$C_{\downarrow}=\sqrt{N}\frac{\sum_{j=r}^{M}j^{-1}}{H_\tau(M)}=\Theta\Big(\sqrt{N}\frac{M^{-1}(M-r)}{\log M}\Big)\overset{M=\Theta(N)}{=}o\Big(\frac{\sqrt{M}}{\log M}\Big).$$

Case $1<\tau<3/2$: Regarding $C_{\updownarrow}$, only $H_{\frac{2\tau}{3}}(M)$ diverges, while the rest of the terms converge. Then, the order of $C_{\updownarrow}$ is determined from $H^{\frac{3}{2}}_{\frac{2\tau}{3}}(M)\sim\left[\frac{M^{1-\frac{2\tau}{3}}-1}{1-\frac{2\tau}{3}}\right]^{\frac{3}{2}}=\Theta\Big(M^{\frac{3}{2}-\tau}\Big)$.
As before, if the condition of Theorem 16 is a strict inequality, $C_{\downarrow}=0$. Otherwise, it is an equality, with $M=\Theta(N)$, thus,

$$C_{\downarrow}=\sqrt{N}\frac{\sum_{j=r}^{M}j^{-\tau}}{H_\tau(M)}\overset{(15)}{=}\Theta\Big(\sqrt{N}M^{-\tau}(M-r)\Big)\overset{M=\Theta(N)}{=}o\Big(M^{\frac{3}{2}-\tau}\Big).$$

Case $\tau=3/2$: $C_{\updownarrow}=\Theta\Big(\log^{\frac{3}{2}}M\Big)$ due to the numerator, all other terms are bounded. If the condition of Theorem 16 is a strict inequality, $C_{\downarrow}=0$. Otherwise, it is an equality, with $M=\Theta(N)$, thus, $C_{\downarrow}=\Theta\Big(\sqrt{N}\frac{M-r}{M^{\frac{3}{2}}}\Big)=\Theta\big(\frac{M-r}{M}\big)=o(1)$. In total, $C=\Theta\Big(\log^{\frac{3}{2}}M\Big)$.
Case $\tau>3/2$: All terms converge in (17), thus $C_{\updownarrow}=O(1)$. If the condition of Theorem 16 is a strict inequality, $C_{\downarrow}=0$. Otherwise, it is an equality, with $M=\Theta\left(N^{\frac{3}{2\tau}}\right)$, thus $C_{\downarrow}=\Theta\Big(\sqrt{N}\frac{M-r}{M^\tau}\Big)=\Theta\Big(\frac{M-r}{M^{\frac{2\tau}{3}}}\Big)=o(1)$. In total, $C=O(1)$. ■

Proof of Theorem 18: In the second case of $KN-M=O(1)$, observe that $KN-M$ is the remaining number of places after putting the M files once in the network. Clearly, $r\leq KN-M$, thus $r=O(1)$. As both r and l are bounded, (26) cannot be true, therefore $\hat{l}=1$. Hence, r is estimated from (23), (24), which, substituting $l\to 1$ yields (34).

For the first part where $KN-M=\omega(1)$, we first note that $r=O(N)$, as $r\leq M$, and $M=O(N)$, due to (1).
Case $\tau<\frac{3}{2}$: From Lemma 11, $l\to\hat{l}=1$. Using this along with (15) in (25), we can estimate r:

$$KN-M+r-1\cong 3(r-1)^{\frac{2\tau}{3}}\frac{r^{1-\frac{2\tau}{3}}-1}{3-2\tau}$$

Observe that assuming $r=O(1)$, the above results become a contradiction, as $NK-M=\omega(1)$, whereas all the other terms are $O(1)$. Therefore, it is $r=\omega(1)$, and (29) follows.
Case $\tau=\frac{3}{2}$: From Lemma 11, $l\to\hat{l}=1$. Working as before, (25) in view of (15) gives that $NK-M+r-1\cong(r-1)\ln r$. Clearly, $r=\omega(1)$, thus, $r\ln r\sim KN-M$.
Case $\tau>\frac{3}{2}$: First, we assume $\hat{l}>1$. Using (22) and (15),

$$K-l+1-\frac{M-r+1}{N}\cong l^{\frac{2\tau}{3}}\frac{l^{1-\frac{2\tau}{3}}-r^{1-\frac{2\tau}{3}}}{\frac{2\tau}{3}-1}\cong\frac{l-l^{\frac{2\tau}{3}}r^{1-\frac{2\tau}{3}}}{\frac{2\tau}{3}-1}\Rightarrow$$

$$K-l+1-\frac{M}{N}+\frac{l}{N^{1-\frac{3}{2\tau}}}\cong 3l\frac{1-\frac{1}{N^{1-\frac{3}{2\tau}}}}{2\tau-3},$$

where in the last step we used (26) to substitute $r\cong lN^{\frac{3}{2\tau}}$. For $N\to\infty$, it is $N^{1-\frac{3}{2\tau}}\to\infty$, and the above becomes

$$\hat{l}\cong\frac{2\tau-3}{2\tau}\left(K+1-\lim\frac{M}{N}\right).$$

Thus, the assumption of $\hat{l}>1$ is correct if $K,\lim M/N$ and τ are such that the second factor of RHS approximately exceeds 1, i.e., $M\overset{\lim}{\leq}\left(K-\frac{3}{2\tau-3}\right)N$.

Then, from (26),

$$r\sim\frac{2\tau-3}{2\tau}\left[(K+1)N^{\frac{3}{2\tau}}-\frac{M}{N^{1-\frac{3}{2\tau}}}\right].$$

Otherwise, $\hat{l}=1$, and r is computed from (25) using (15)

$$NK-M+r-1\cong 3(r-1)^{\frac{2\tau}{3}}\frac{1-r^{1-\frac{2\tau}{3}}}{2\tau-3}.$$

As $N\to\infty$, it follows that $r\sim\left[\frac{2\tau-3}{3}(KN-M)\right]^{\frac{3}{2\tau}}$. ■

Proof of Theorem 19: First note that from Theorem 18, for all τ, it is $K_{\updownarrow}=\Theta\left(\frac{KN-M+\hat{r}-1}{N}\right)=\Theta(1)$ (using $M\overset{\lim}{<}KN$).

In the cases of $\tau<\frac{3}{2}$, $\mathcal{M}_{\downarrow}\neq\emptyset$ entails $M\overset{\lim}{>}K\left(1-\frac{2\tau}{3}\right)N$ (Theorem 16). It is also $M\overset{\lim}{<}KN$, thus $M=\Theta(N)$.

Furthermore, from Theorem 18 and $M\overset{\lim}{<}KN$, it is

$$r\sim\frac{3-2\tau}{2\tau}(KN-M)+1\overset{M\overset{\lim}{<}KN}{=}\Theta(N),\ \text{and, moreover,}$$

$$r\overset{\lim}{<}\frac{3-2\tau}{2\tau}\,\frac{2\tau}{3}KN=\left(1-\frac{2\tau}{3}\right)KN\overset{\lim}{<}M. \tag{42}$$

Then, we compute the link rate as follows:
Case $\tau<1$: Using Lemma 13, it is $C_{\downarrow}=\Theta\left(\sqrt{N}\right)$. Invoking Lemma 12, too, we get that $C=\Theta\Big(\sqrt{N}\Big)=\Theta\Big(\sqrt{M}\Big)$.
Case $\tau=1$: $C_{\downarrow}=\sqrt{N}\frac{H_1(M)-H_1(r)}{H_1(M)}\overset{(15)}{\sim}\sqrt{N}\frac{\ln\frac{M}{r}}{\ln M}=\Theta\Big(\frac{\sqrt{M}}{\log M}\Big)$,

using $r=\Theta(N)=\Theta(M)$. Similarly, as $l\to 1$, $C_{\updownarrow}=\Theta\left(\frac{H^{\frac{3}{2}}_{\frac{2}{3}}(r-1)}{\sqrt{K_{\updownarrow}}\ H_1(M)}\right)=\Theta\Big(\frac{\sqrt{N}}{\log N}\Big)$. In total, $C=\Theta\Big(\frac{\sqrt{M}}{\log M}\Big)$.
Case $1<\tau<3/2$: Using $r=\Theta(N)=\Theta(M)$,

$$C_{\downarrow}=\sqrt{N}\ \frac{H_\tau(M)-H_\tau(r)}{H_\tau(M)}\overset{(15)}{\sim}\frac{\sqrt{M}}{r^{\tau-1}}\Big[1-\Big(\frac{r}{M}\Big)^{\tau-1}\Big],$$

which is $C_{\downarrow}=O\left(M^{\frac{3}{2}-\tau}\right)$ from (42). Last, $l\to 1$ implies that $C_{\updownarrow}\sim\frac{H^{\frac{3}{2}}_{\frac{2\tau}{3}}(r-1)}{\sqrt{K_{\updownarrow}}\ H_\tau(M)}=\Theta\left(M^{\frac{3}{2}-\tau}\right)$. In total, $C=\Theta\left(M^{\frac{3}{2}-\tau}\right)$.
Case $\tau=3/2$: Now, it has to be $M\ln M\overset{\lim}{>}KN$, which also implies that $M\log M=\Omega(N)$. From Theorem 18, we have that $r\ln r\sim KN-M$. This means that $r\log r=\Theta(N)$ in view of $M\overset{\lim}{<}KN$, and thus $r=o(N)$.

Moreover, comparing $M\ln M$ and $r\ln r$ in the above formulas, it has to be $r\overset{\lim}{<}M$. The latter implies that there exists a $0<k<1$ such that eventually $\frac{r}{M}\leq k$. Using then (18),

$$C_{\downarrow}\overset{(15)}{=}\Theta\Big(N^{\frac{1}{2}}\Big[\frac{1}{r^{\frac{1}{2}}}-\frac{1}{M^{\frac{1}{2}}}\Big]\Big)\overset{(\frac{r}{M})^{\frac{1}{2}}\leq\sqrt{k}}{=}\Theta\left(\sqrt{\frac{N}{r}}\right)$$

$$= \Theta\left(\sqrt{\frac{N}{KN-M}\log r}\right) \overset{NK-M=\Theta(N)}{=} \Theta\left(\sqrt{\log r}\right).$$

Moreover, as $l \to 1$, $C_{\circ} = \Theta\left(\frac{H_1^{\frac{3}{2}}(r)}{\sqrt{K_{\circ}}\ H_{\frac{3}{2}}(M)}\right) = \Theta\left((\log r)^{\frac{3}{2}}\right)$.

Thus, in total $C = \Theta\left((\log r)^{\frac{3}{2}}\right)$.

Case $\tau > 3/2$: it is $r = \Theta\left(N^{\frac{3}{2\tau}}\right)$ due to $M \overset{\lim}{<} KN$. Moreover, for $\mathcal{M}_{\downarrow} \neq \emptyset$, it has to be $M = \Omega\left(N^{\frac{3}{2\tau}}\right)$. Then,

$$C_{\downarrow} = \sqrt{N}\frac{H_\tau(M) - H_\tau(r-1)}{H_\tau(M)} \overset{((15))}{=} O\left(N^{\frac{1}{2}}r^{1-\tau}\right)$$
$$= O\left(N^{\frac{1}{2}+\frac{3}{2\tau}(1-\tau)}\right) = O\left(N^{\frac{3}{2\tau}-1}\right) = O(1).$$

Last, $C_{\circ} = \Theta(1)$ (all terms converge). Thus, $C = \Theta(1)$. ∎

Proof of Theorem 20: In all the cases, we know that $r \leq KN - M + 1$, as $KN - M$ is the number of spaces left for duplicate copies after all M files are stored once. Hence, $r = O(KN - M) = o(N) = o(M)$. Moreover, as before, in all cases, $K_{\circ} = \Theta\left(\frac{KN-M+r-1}{N}\right) = \Theta\left(K - \frac{M}{N}\right)$.

Case $\tau \leq 1$: From Lemma 13, $r = \omega(M)$ implies that $C_{\downarrow} = \Theta\left(\sqrt{N}\right)$. Hence, invoking Lemma 12, $C \overset{M=\Theta(N)}{=} \Theta\left(\sqrt{M}\right)$.

For the rest of the cases with $\tau > 1$, it is $r = o(M)$, therefore, from Lemma 13, we get that $C_{\downarrow} = \Theta\left(\frac{\sqrt{N}}{r^{\tau-1}}\right)$.

Case $1 < \tau < 3/2$: Using $r = \Theta(KN-M)$ from Theorem 18, $C_{\downarrow} = \Theta\left(\frac{\sqrt{N}}{(KN-M)^{\tau-1}}\right)$. On the other hand, $l \to 1$, and thus $C_{\circ} = \frac{H_{\frac{2\tau}{3}}^{\frac{3}{2}}(r-1)}{\sqrt{K_{\circ}}H_\tau(M)} = \Theta\left(\sqrt{\frac{N}{KN-M}}r^{\frac{3}{2}-\tau}\right) = O\left(\frac{\sqrt{N}}{(KN-M)^{\tau-1}}\right)$. In total, $C \overset{M=\Theta(N)}{=} \Theta\left(\frac{\sqrt{M}}{(KN-M)^{\tau-1}}\right)$.

Case $\tau = 3/2$: From the above, $C_{\downarrow} = \Theta\left(\sqrt{\frac{N}{r}}\right)$. Moreover,

$$C_{\circ} = \frac{H_1^{\frac{3}{2}}(r-1)}{\sqrt{K_{\circ}}H_{\frac{3}{2}}(M)} = \Theta\left(\sqrt{\frac{N}{KN-M}}\log^{\frac{3}{2}} r\right).$$

However, $\frac{1}{r} = \frac{\log r}{r\log r} = \Theta\left(\frac{\log r}{KN-M}\right) = o\left(\frac{\log^3 r}{KN-M}\right)$, thus $C_{\downarrow} = o(C_{\circ})$. In total, $C \overset{M=\Theta(N)}{=} \Theta\left(\sqrt{\frac{M}{KN-M}}\log^{\frac{3}{2}} r\right)$.

Case $\tau > 3/2$: From Theorem 18, $r = \Theta\left((KN-M)^{\frac{3}{2\tau}}\right) = o(KN - M) = o(M)$. Thus, $C_{\downarrow} = \Theta\left(\frac{\sqrt{N}}{(KN-M)^{\frac{3}{2}\ \frac{\tau-1}{\tau}}}\right)$. Moreover, $C_{\circ} = \Theta\left(K_{\circ}^{-\frac{1}{2}}\right) = \Theta\left(\sqrt{\frac{N}{KN-M}}\right)$ (the H-terms converge). As $\frac{3(\tau-1)}{2\tau} > \frac{1}{2}$, it is $C \overset{M=\Theta(N)}{=} \Theta\left(\sqrt{\frac{M}{KN-M}}\right)$. ∎

References

[1] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, and R. L. Braynard, "Networking named content," in *Proc. of the 5th international conference on Emerging networking experiments and technologies (CoNEXT '09)*, Rome, Italy, Dec. 2009, pp. 1–12.

[2] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Trans. Inf. Theory*, vol. 46, pp. 388–404, Mar. 2000.

[3] A. Özgür, O. Lévêque, and D. Tse, "Hierarchical cooperation achieves optimal capacity scaling in ad hoc networks," *IEEE Trans. Inf. Theory*, vol. 53, pp. 3549–3572, Oct. 2007.

[4] S. Toumpis, "Asymptotic capacity bounds for wireless networks with non-uniform traffic patterns," *IEEE Trans. Wireless Commun.*, vol. 7, pp. 2231–2242, Jun. 2008.

[5] A. Zemlianov and G. de Veciana, "Capacity of ad hoc wireless networks with infrastructure support," *IEEE J. Sel. Areas Commun.*, vol. 23, pp. 657–667, Mar. 2005.

[6] M. Franceschetti, M. D. Migliore, and P. Minero, "The capacity of wireless networks: information-theoretic and physical limits," *IEEE Trans. Inf. Theory*, vol. 55, pp. 3413–3424, Aug. 2009.

[7] S. Gitzenis, G. S. Paschos, and L. Tassiulas, "Asymptotic laws for content replication and delivery in wireless networks," in *The 31st Annual IEEE International Conference on Computer Communications (INFOCOM'2012)*, Orlando, Florida, USA, Mar. 2012.

[8] S. Jin and L. Wang, "Content and service replication strategies in multi-hop wireless mesh networks," in *MSWiM '05: Proceedings of the 8th ACM international symposium on Modeling, analysis and simulation of wireless and mobile systems*, Montréal, Quebec, Canada, Oct. 2005, pp. 79–86.

[9] V. Sourlas, G. S. Paschos, P. Flegkas, and L. Tassiulas, "Mobility support through caching in content-based publish-subscribe networks," in *Proc. of 5th International workshop on content delivery networks (CDN 2010)*, Melbourne, Australia, May 2010, pp. 715–720.

[10] J. Zhao, P. Zhang, G. Cao, and C. R. Das, "Cooperative caching in wireless p2p networks: Design, implementation, and evaluation," *IEEE Trans. Parallel Distrib. Syst.*, vol. 21, pp. 229–241, Feb. 2010.

[11] U. Niesen, D. Shah, and G. Wornell, "Caching in wireless networks," in *IEEE International Symposium on Information Theory*, Seoul, Korea, Jun. 2009, pp. 2111–2115.

[12] J. Silvester and L. Kleinrock, "On the capacity of multihop slotted aloha networks with regular structure," *IEEE Trans. Commun.*, vol. 31, pp. 974–982, Aug. 1983.

[13] M. E. J. Newman, "Power laws, pareto distributions and Zipf's law," *Contemporary Physics*, vol. 46, pp. 323–351, Sep./Oct. 2005.

[14] L. A. Adamic and B. A. Huberman, "Zipf's law and the Internet," *Glottometrics*, vol. 3, pp. 143–150, 2002.

[15] J. Chu, K. Labonte, and B. N. Levine, "Availability and popularity measurements of peer-to-peer file systems," in *Proceedings of SPIE*, Boston, MA, USA, Jul. 2002.

[16] T. Yamakami, "A Zipf-like distribution of popularity and hits in the mobile web pages with short life time," in *Proc. of Parallel and Distributed Computing, Applications and Technologies, PDCAT '06*, Taipei, ROC, Dec. 2006, pp. 240–243.

[17] L. Breslau, P. Cue, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. of INFOCOM*, New York, NY, USA, Mar. 1999, pp. 126–134.

[18] C. R. Cunha, A. Bestavros, and M. E. Crovella, "Characteristics of WWW Client-based Traces," in *View on NCSTRL*, Boston University, MA, USA, Jul. 1995.

[19] M. Franceschetti and R. Meester, *Random Networks for Communication*. New York, NY, USA: Cambridge University Press, Series: Cambridge Series in Statistical and Probabilistic Mathematics (No. 24), 2007.

[20] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.